# A skew-$t$-normal multi-level reduced-rank functional PCA model with applications to replicated 'omics time series data sets

**Maurice Berk**

*Section of Paediatrics*
*Department of Medicine*
*Imperial College London*
*Norfolk Place*
*London*
*W2 1PG*
*e-mail:* maurice.berk01@imperial.ac.uk

**and**

**Giovanni Montana***

*Statistics Section*
*Department of Mathematics*
*Imperial College London*
*Huxley Building*
*London*
*SW7 2AZ*
*e-mail:* giovanni.montana@imperial.ac.uk

**Abstract:** A powerful study design in the fields of genomics and metabolomics is the 'replicated time course experiment' where individual time series are observed for a sample of biological units, such as human patients, termed replicates. Standard practice for analysing these data sets is to fit each variable (e.g. gene transcript) independently with a functional mixed-effects model to account for between-replicate variance. However, such an independence assumption is biologically implausible given that the variables are known to be highly correlated.

In this article we present a skew-$t$-normal multi-level reduced-rank functional principal components analysis (FPCA) model for simultaneously modelling the between-variable and between-replicate variance. The reduced-rank FPCA model is computationally efficient and, analogously with a standard PCA for vectorial data, provides a low dimensional representation that can be used to identify the major patterns of temporal variation. Using an example case study exploring the genetic response to BCG infection we demonstrate that these low dimensional representations are eminently biologically interpretable. We also show using a simulation study that modelling all variables simultaneously greatly reduces the estimation error compared to the independence assumption.

*We are grateful to Cheryl Hemingway and Timothy Ebbels for providing access to the example data sets used in this article

## 1. Introduction

Genomics and metabolomics are two examples of a broader range of 'omics domains, each of which characterises a biological organism at a different level of biomolecular organisation. A powerful study design within these fields is the 'replicated time course experiment', in which a sample of biological units such as human patients or laboratory rats, termed 'replicates', is studied over time in order to infer the temporal behaviour of the population as a whole. As biological processes are inherently dynamic, these time series experiments provide greater insight than static analyses. Data arising from these experiments presents some unique challenges compared to more traditional time series analysis application domains. In particular, the time series are very short with 5 to 10 time points being typical. This is due to the expense involved in collecting observations, not purely in monetary terms but also due to ethical concerns about obtaining the biological samples, and the laboratory time needed to conduct the assays. The number of replicates are equally small. Due to the specifics of the assaying technologies utilised, the observations are often collected with a great deal of noise, and some may simply be missing. Missing data may also arise by design, especially when it is necessary to sacrifice replicates in animal studies in order to obtain the biological samples. Experimental design may also lead to the time points being irregularly spaced, in an attempt to exploit *a priori* knowledge about the temporal behaviour. Finally, the data is high dimensional with tens of thousands of variables (e.g. gene transcripts) under study simultaneously.

In order to deal with these issues, functional data analysis (FDA) (Ramsay and Silverman, 2005) has become a popular modelling choice in the field of genomics time series data analysis. In FDA, we assume that our observations are noisy realisations of an underlying smooth function of time (or, analogously, a curve) which is to be estimated. After estimation, this function is then treated as the fundamental unit of data in any subsequent analysis, such as clustering or network inference. Within the field of genomics, FDA approaches have been proposed for clustering unreplicated data sets (Ma et al., 2006), detecting significant genes in multi-sample replicated data sets (Storey et al., 2005) and detecting significant genes in cross-sectional studies (Angelini, Canditiis and Pensky, 2009). We ourselves have demonstrated that such methodology is equally well-suited to the field of metabolomics (Berk, Ebbels and Montana, 2011; Montana, Berk and Ebbels, 2011).

Despite the seeming essentiality of replication, the replicated time course study design is rare, possibly due to a lack of appreciation that without replication inference is restricted to the single sample under study alone, or perhaps limited by the few adequate modelling choices available. This limited range of statistical methodology has no doubt arisen from the complexity involved in simultaneously modelling the covariance between the variables and, for a given variable, between the replicates. This complexity is exacerbated by the high dimensionality of the data which incurs a significant computational cost.

There are only a handful of approaches specifically designed for replicated genomics time series data sets that we are aware of. The first of these, proposed

by Tai and Speed (2009), does not truly account for time as a quantitative variable in the sense that the results of an analysis would be the same if the time points were to be permuted. Furthermore, it cannot handle missing data without resorting to undesirable imputation procedures. The other two approaches, the functional mixed-effects model of Storey et al. (2005), and the functional principal components analysis (FPCA) method proposed by Liu and Yang (2009) both opt to avoid the complexity of simultaneously modelling both levels of covariance by instead modelling each variable independently. In the case of Storey et al. (2005), each variable is summarised with a mean curve across all replicates. The replicate effects are treated as scalar shifts from this mean curve, so that each replicate exhibits exactly the same temporal profile. We have previously extended this approach to allow for more realistic heterogeneous replicate behaviour by treating the replicate effects themselves as curves (Berk et al., 2010). Under this model, each variable can be summarised with a mean curve and covariance surface which describes the replicate heterogeneity. Similarly, Liu and Yang (2009) use the 'principal analysis through conditional expectation' (PACE) method of Yao, Müller and Wang (2005), also summarising each variable with a mean curve, but this time with an eigen-decomposition of the covariance surface.

All of the methods mentioned above rely on the assumption of independence between variables. It is clear to understand the motivation for this independence assumption as, while it may be biologically unjustified, it greatly simplifies the analysis. However, given that there have been no proposed alternatives that do account for both levels of covariance, it has not been possible to ascertain the true impact of this simplification, for example in terms of estimation error. However, it is reasonable to assume, given the few observations available for each variable, that the effect is significant.

There have been two methods proposed for accounting for multiple levels of covariance outside of 'omics application domains. The first of these, introduced by Di et al. (2009), suggests to estimate the covariance surface at each level using the method of moments, which is then smoothed using thin-plate spline-smoothing. For dimensionality reduction, the resulting surfaces are subject to eigen-decompositions. We remain sceptical, however, of the applicability of such an approach to 'omics data sets where the tiny number of time points and few replicates raises significant concerns as to the ability of the method of moments to adequately estimate the covariance surfaces. In contrast, Di et al. (2009) demonstrate their method on a data set with over $3,000$ replicates and 960 time points.

The other proposed approach is the multi-level reduced-rank FPCA model of Zhou et al. (2010), extending the single-level reduced-rank FPCA model of James, Hastie and Sugar (2000). This is similar to the model that we introduce in this article, however the key difference is that they assume the principal component loadings at each level are normally distributed. We will demonstrate here that such an assumption is untenable at the variable level for 'omics data sets, where the small number of variables with significantly time varying profiles, in conjunction with the high dimensionaliy of the data, leads to distributions

which exhibit a high degree of kurtosis and which may be skewed. In order to address this issue we propose a multi-level reduced-rank FPCA model in which the variable level loadings follow a skew-$t$-normal distribution, which is a flexible four parameter distribution allowing for both heavy tails and skewness.

## 2. Methods

### *2.1. A multi-level reduced-rank FPCA model*

We assume that the observation, such as gene expression level or NMR spectrum intensity, at time $t$ on replicate $j$ for variable $i$, $y_{ij}(t)$, is described by the following functional mixed-effects model:

$$y_{ij}(t) = \mu(t) + f_i(t) + g_{ij}(t) + \epsilon_i(t) \tag{1}$$

where $\mu(t)$ is the 'grand mean' across all variables and $f_i(t)$ is the offset from the grand mean for variable $i$, so that $\mu(t) + f_i(t)$ represents the mean function for variable $i$; $g_{ij}(t)$ is the replicate offset from the variable mean for replicate $j$, specific to variable $i$; $\epsilon_i(t)$ is an error term, specific to variable $i$. Note that if the replicate effect $g_{ij}(t)$ was not variable specific then the subscript $i$ could be dropped so that the same $g_j(t)$ term was shared across all variables. However, both intuition and real data sets support the idea of separate replicate effects for each variable. We would expect in a genomics experiment, for instance, that certain gene transcripts display a homogeneous response across all replicates while others are much more heterogeneous, and it could well be the case that these differences lead to exactly the biological effect we are seeking to identify. Even when two transcripts both display heterogeneity between the replicates, the exact nature of that variation is likely to be transcript-specific. In Supplementary Figures 1 and 2 we give raw data for two transcripts from our example data set that illustrate these points.

As in James, Hastie and Sugar (2000), we can simultaneously achieve computational efficiency, parsimony and dimensionality reduction by replacing $f_i(t)$ and $g_{ij}(t)$ in (1) with their Karhunen-Loève decompositons, yielding

$$y_{ij}(t) = \mu(t) + \sum_{k=1}^{\infty} \zeta_k(t)\alpha_{ik} + \sum_{l=1}^{\infty} \eta_{il}(t)\beta_{ijl} + \epsilon_{ij}(t)$$

where $\zeta_k(t)$ is the $k$-th principal component function at the variable level, $\alpha_{ik}$ is variable $i$'s loading on the $k$-th principal component function, $\eta_{il}(t)$ is the $l$-th principal component function at the replicate level, specific to variable $i$ and $\beta_{ijl}$ is replicate $j$'s loading on the $l$-th principal component function specific to variable $i$. Truncating the decompositions at the $K$-th and $L_i$-th component for the variable and replicate level respectively gives

$$y_{ij}(t) = \mu(t) + \sum_{k=1}^{K} \zeta_k(t)\alpha_{ik} + \sum_{l=1}^{L_i} \eta_{il}(t)\beta_{ijl} + \epsilon_{ij}(t)$$

Note that the number of principal components retained at the replicate level, $L_i$, is variable specific as indicated by the subscript. The optimal number of principal components to be retained at this level will differ between variables depending upon the amount of between-replicate heterogeneity displayed. Representing the functions $\mu(t)$, $\zeta_k(t)$ and $\eta_{il}(t)$ using an appropriately orthogonalised $p$-dimensional B-spline basis (Zhou, Huang and Carroll, 2008) and collecting all $N_{ij}$ observations on replicate $j$ for variable $i$ in the vector $\boldsymbol{y}_{ij}$ yields

$$\boldsymbol{y}_{ij} = \boldsymbol{B}_{ij}\boldsymbol{\theta}_\mu + \sum_{k=1}^{K} \boldsymbol{B}_{ij}\boldsymbol{\theta}_{\alpha_k}\alpha_{ik} + \sum_{l=1}^{L_i} \boldsymbol{B}_{ij}\boldsymbol{\theta}_{\beta_{il}}\beta_{ijl} + \boldsymbol{\epsilon}_{ij}$$

where $\boldsymbol{B}_{ij}$ is the $N_{ij} \times p$ B-spline basis matrix that has been transformed such that $(L/g)\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}$ where $\boldsymbol{B}$ is the basis evaluated on a fine grid of points covering the range of the time course, $L$ is the length of this fine grid, $g$ is the distance between successive grid points, $\boldsymbol{\theta}_\mu$ is a $p$-length vector of fitted spline coefficients for the grand mean function $\mu(t)$, $\boldsymbol{\theta}_{\alpha_k}$ is a $p$-length vector of fitted spline coefficients for the $k$-th principal component function at the variable level and $\boldsymbol{\theta}_{\beta_{il}}$ is a $p$-length vector of fitted spline coefficients for the $l$-th principal component function at the replicate level. The transformation of $\boldsymbol{B}$ is required to enforce the FPCA orthogonality constraint that

$$\int_t \zeta_k(t)\zeta_{k'}(t) = \begin{cases} 1 & k = k' \\ 0 & k \neq k' \end{cases}$$

and similarly for $\eta_{il}(t)$.

Defining the $p \times K$ matrix $\boldsymbol{\Theta}_\alpha = [\boldsymbol{\theta}_{\alpha_1} \cdots \boldsymbol{\theta}_{\alpha_K}]$ and the $p \times L_i$ matrix $\boldsymbol{\Theta}_{\beta_{ij}} = [\boldsymbol{\theta}_{\beta_{i1}} \cdots \boldsymbol{\theta}_{\beta_{iL_i}}]$ allows the summations to be simplified using matrix algebra as

$$\boldsymbol{y}_{ij} = \boldsymbol{B}_{ij}\boldsymbol{\theta}_\mu + \boldsymbol{B}_{ij}\boldsymbol{\Theta}_{\alpha_k}\boldsymbol{\alpha}_i + \boldsymbol{B}_{ij}\boldsymbol{\Theta}_{\beta_i}\boldsymbol{\beta}_{ij} + \boldsymbol{\epsilon}_{ij} \tag{2}$$

where $\boldsymbol{\alpha}_i = [\alpha_{i1} \cdots \alpha_{iK}^T]$ is the $K$-length vector formed by collecting all of the $\alpha_{ik}$ terms together and similarly for $\boldsymbol{\beta}_{ij}$. By collecting the observations on all replicates $j = 1, \cdots, n_i$ for variable $i$, in the vector $\boldsymbol{y}_i = [\boldsymbol{y}_{i1} \cdots \boldsymbol{y}_{in_i}]^T$, we can write

$$\boldsymbol{y}_i = \boldsymbol{B}_i\boldsymbol{\theta}_\mu + \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i + \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{B}_i = [\boldsymbol{B}_{i1}^T \cdots \boldsymbol{B}_{in_i}^T]^T$ is the $N_i \times p$ basis matrix formed by stacking each $\boldsymbol{B}_{ij}$ matrix on top of one another. There are $n_i$ such matrices, each with $N_{ij}$ rows and so the matrix $\boldsymbol{B}_i$ has $N_i$ rows where $N_i = \sum_{i=1}^{n_i} N_{ij}$ is the total number of observations on variable $i$ across all replicates. In contrast, the matrix $\widetilde{\boldsymbol{B}_i} = \text{diag}(\boldsymbol{B}_{i1}, \cdots, \boldsymbol{B}_{in_i})$ is a block diagonal matrix of dimension $N_i \times (n_i p)$ where the blocks correspond to the $\boldsymbol{B}_{ij}$ matrices. Similarly, $\widetilde{\boldsymbol{\Theta}_{\beta_i}} = \text{diag}(\boldsymbol{\Theta}_{\beta_i}, \cdots, \boldsymbol{\Theta}_{\beta_i})$ is a block diagonal matrix of dimension $(n_i L_i) \times (n_i L_i)$ where each block is identical and equal to $\boldsymbol{\Theta}_{\beta_i}$. Finally, $\boldsymbol{\beta}_i = [\boldsymbol{\beta}_{i1} \cdots \boldsymbol{\beta}_{in_i}]^T$ and $\boldsymbol{\epsilon}_i = [\boldsymbol{\epsilon}_{i1} \cdots \boldsymbol{\epsilon}_{in_i}]^T$.

Standard practice would be to assume that $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are all independently multivariate normally distributed with zero mean and covariance matrices $\boldsymbol{D}_\alpha$, $\boldsymbol{D}_{\beta_i}$ and $\sigma_i^2\boldsymbol{I}$ respectively. Under these assumptions, $\boldsymbol{y}_i$ is marginally

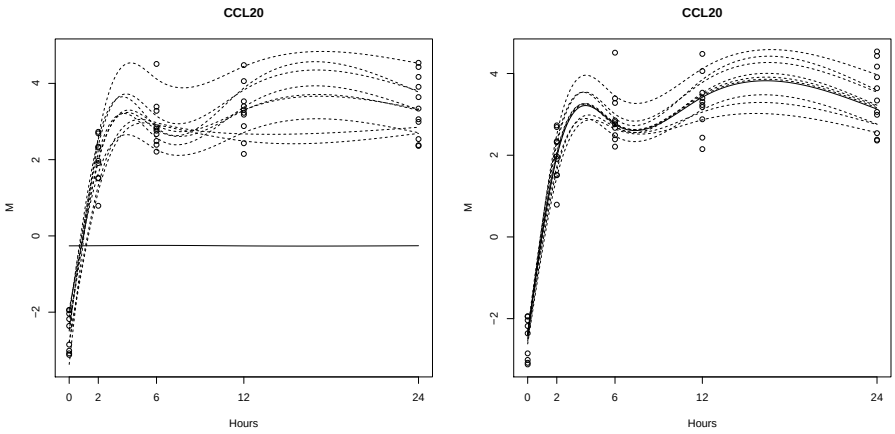



FIG 1. *Left: An example of the poor model fits obtained under the Gaussian reduced-rank multi-level FPCA model with real data. This example gene transcript is typical of variables across a range of data sets, where, while the replicate curves, indicated by the dashed lines, look good and closely map the underlying observations, the mean curve, given by the solid line, seems to be biologically implausible. Right: This poor fit is fixed when we instead adopt our proposed skew-t-normal multi-level reduced-rank FPCA model.*

multivariate normal and the model parameters can be estimated by treating the principal component loadings $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_{ij}$ as missing data and employing the EM algorithm. Technical details for this approach can be found in the Supplementary Material.

### *2.2. A skew-t-normal multi-level reduced-rank FPCA model*

We discovered upon fitting the multi-level reduced-rank FPCA model with the normal assumption to real data that the fits were biologically implausible, an example of which is given on the left hand side of Figure 1. As can be clearly seen, the mean curve for this gene transcript, indicated by the solid line, does an unusually poor job of describing the underlying observations, being flat over the entire range of the time course. However, the replicate-level curves, indicated by the dashed lines, follow the observations much more closely. Upon further investigation, it became clear that this was due to an extreme departure from normality for the principal component loadings at the variable level, as can be seen in histograms of the initalised loadings for two example data sets given in Supplementary Figures 3 and 4, which exhibit heavy tails and varying degrees of skewness. To deal with this issue, we propose to instead adopt the assumption that the variable level loadings follow a skew-$t$-normal distribution, which is flexible enough to account for the heterogenous departures from normality.

Formally, we assume that:

$$\begin{gathered}\alpha_{ik} \overset{\text{i.i.d.}}{\sim} StN(\xi_{\alpha_k}, \sigma^2_{\alpha_k}, \lambda_{\alpha_k}, \nu_{\alpha_k}) \\ \alpha_{ik} \perp \alpha_{ik'}, k \neq k' \quad \alpha_{ik} \perp \alpha_{i'k}, i \neq i' \\ E[\alpha_{ik}] = 0\end{gathered} \tag{3}$$

while retaining the assumption that $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are multivariate normally distributed. $z \sim StN(\xi, \sigma^2, \lambda, \nu)$ denotes that the random variable $z$ follows a skew-$t$-normal distribution (Gómez, Venegas and Bolfarine, 2007) where $\xi$ is a location parameter, $\sigma^2$ is a scale parameter, $\lambda$ is a skewness parameter and $\nu$ is the degrees of freedom controlling the kurtosis. The density of $z$ is given by

$$f(z|\xi, \sigma^2, \lambda, \nu) = 2t_\nu(z; \xi, \sigma^2)\Phi\left(\frac{z-\xi}{\sigma}\lambda\right) \tag{4}$$

where $t_\nu(z; \xi, \sigma^2)$ denotes the Student-$t$ density with $\nu$ degrees of freedom, location parameter $\xi$ and scale parameter $\sigma^2$, and $\Phi$ denotes the normal cumulative distribution function. We note here in passing the closely related skew-$t$ distribution of Azzalini and Capitanio (2003) which is identical in form to (4) except that the the normal cumulative distribution function, which controls the skewness of the density, is replaced by the Student-$t$ cumulative distribution function, and depends not just on the skewness parameter $\lambda$ but also the degrees of freedom $\nu$. As Ho and Lin (2010) point out, evaluating the skew-$t$-normal density is therefore computationally simpler and the decoupling of the skewness from the degrees of freedom parameter is more conceptually sound.

Note that we have chosen not to use a multivariate skew-$t$-normal density for the variable-level loadings as this would require a single degrees of freedom parameter to be shared across all components, and we assume that they are independent anyway for the purposes of identifiability. Furthermore, we retain the original assumption that the replicate-level loadings follow a multivariate normal distribution. Judging from Figure 1, which is typical of other variables across multiple data sets, the replicate-level curves remain plausible under this assumption.

As in the Gaussian case, we can estimate the parameters under the assumptions given in (3) by treating the principal component loadings as missing data and employing the EM algorithm. The maximum likelihood estimators of $\boldsymbol{\theta}_\mu$, $\boldsymbol{\Theta}_\alpha$, $\boldsymbol{\Theta}_{\beta_i}$, $\boldsymbol{D}_{\beta_i}$ and $\sigma_i^2$ can be derived analytically as illustrated in the Supplementary Material. For the parameters of the skew-$t$-normal distributions, these can be estimated using Newton-Raphson (NR) as in Gómez, Venegas and Bolfarine (2007) or the EM algorithm as in Ho and Lin (2010). However, the NR approach relies on numerical integration and the EM algorithm is known to be slow to converge; alternatively, we have found that a simplex optimisation (Nelder and Mead, 1965) works very well in practice.

The conditional expectations that need to be calculated at the E-step of the EM algorithm based on these MLE estimators are summarised in Supplementary Table 2. However, expressions for these are challenging to obtain given that

under the distributional assumptions, the marginal density of $\boldsymbol{y}_i$, as a sum of skew-$t$-normal and normal distributed random variables, follows no known form. In these circumstances we can instead turn to the *Monte Carlo* (MC) EM-algorithm (Wei and Tanner, 1990) which replaces the calculation of intractable expectations at the E-step with approximations based on averaging random samples drawn from the target density. As the target density is itself unknown, it is necessary to resort to stochastic simulation algorithms.

Given that the skew-$t$-normal distribution admits a convenient hierarchical representation, we suggest the use of the Gibbs sampler for drawing samples from the joint conditional distribution $f(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_{ij}|\boldsymbol{y}_i)$. Specifically, following Ho and Lin (2010), if $\tau_{ik} \sim \Gamma(\nu_{\alpha_k}/2, \nu_{\alpha_k}/2)$ then

$$\begin{aligned} \gamma_{ik}|\tau_{ik} &\sim TN\left(0, \frac{\tau_{ik}+\lambda_{\alpha_k}^2}{\tau_{ik}}; (0,\infty)\right) \\ \alpha_{ik}|\gamma_{ik}, \tau_{ik} &\sim N\left(\xi_{\alpha_k} + \frac{\sigma_{\alpha_k}\lambda_{\alpha_k}}{\tau_{ik}+\lambda_{\alpha_k}^2}\gamma_{\alpha_k}, \frac{\sigma_{\alpha_k}^2}{\tau_{ik}+\lambda_{\alpha_k}^2}\right) \end{aligned} \tag{5}$$

where $TN(\mu, \sigma^2; (a,b))$ denotes the truncated normal distribution lying within the interval $(a,b)$. It can then be shown (see Supplementary Material) that

$$\gamma_{ik}|\alpha_{ik} \sim TN((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}, 1; (0,\infty)) \tag{6}$$

$$\tau_{ik}|\alpha_{ik} \sim \Gamma\left(\frac{\nu_{\alpha_k}+1}{2}, \frac{\nu_{\alpha_k}+(\alpha_{ik}-\xi_{\alpha_k})^2/\sigma_{\alpha_k}^2}{2}\right) \tag{7}$$

Therefore we introduce additional latent variables, $\boldsymbol{\tau}_i = [\tau_{i1}\cdots\tau_{iK}]^T$ and $\boldsymbol{\gamma}_i = [\gamma_{i1}\cdots\gamma_{iK}]^T$, and the target density for the Monte Carlo E-step becomes the joint conditional distribution $f(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i|\boldsymbol{y}_i)$ whose form is still unknown. However, the conditionals $f(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i)$, $f(\boldsymbol{\tau}_i|\boldsymbol{y}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i)$ and $f(\boldsymbol{\gamma}_i|\boldsymbol{y}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\tau}_i)$ follow known distributions that are easy to sample from and hence the Gibbs sampler can be used to efficiently generate samples that are approximately distributed according to the target full joint conditional density.

Starting with $f(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i)$ note that from (5), conditional on $\tau_{ik}$ and $\gamma_{ik}$, $\alpha_{ik}$ is normally distributed with mean $\mu_{\alpha_k} = \xi_{\alpha_k} + \frac{\sigma_{\alpha_k}\lambda_{\alpha_k}}{\tau_{ik}+\lambda_{\alpha_k}^2}\gamma_{\alpha_k}$ and variance $v_{\alpha_k} = \frac{\sigma_{\alpha_k}^2}{\tau_{ik}+\lambda_{\alpha_k}^2}$. Hence we can write

$$\begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \\ \boldsymbol{y}_i \end{bmatrix} \Bigg| \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i \sim MVN\left(\begin{bmatrix} \boldsymbol{\mu}_\alpha \\ \boldsymbol{0} \\ \boldsymbol{B}_i\boldsymbol{\theta}_\mu + \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{\mu}_\alpha \end{bmatrix}, \begin{bmatrix} \mathrm{diag}(\boldsymbol{v}_\alpha) & \boldsymbol{0} & \mathrm{diag}(\boldsymbol{v}_\alpha)\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T \\ \boldsymbol{0} & \widetilde{\boldsymbol{D}_{\beta_i}} & \widetilde{\boldsymbol{D}_{\beta_i}}\widetilde{\boldsymbol{\Theta}_{\beta_i}^T}\widetilde{\boldsymbol{B}_i^T} \\ \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\mathrm{diag}(\boldsymbol{v}_\alpha) & \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widetilde{\boldsymbol{D}_{\beta_i}} & \boldsymbol{V}_{y_i|\tau_i,\gamma_i} \end{bmatrix}\right)$$

where $\boldsymbol{\mu}_\alpha = [\mu_{\alpha_1} \cdots \mu_{\alpha_K}]^T$ and $\boldsymbol{v}_\alpha = [v_{\alpha_1} \cdots v_{\alpha_K}]^T$, $\boldsymbol{V}_{y_i|\tau_i,\gamma_i} = \sigma_i^2 \boldsymbol{I}_{N_i \times N_i} + \boldsymbol{B}_i \boldsymbol{\Theta}_\alpha \text{diag}(\boldsymbol{v}_\alpha) \boldsymbol{\Theta}_\alpha^T \boldsymbol{B}_i^T + \widetilde{\boldsymbol{B}_i} \widetilde{\boldsymbol{\Theta}_{\beta_i}} \widetilde{\boldsymbol{D}_{\beta_i}} \widetilde{\boldsymbol{\Theta}_{\beta_i}^T} \widetilde{\boldsymbol{B}_i^T}$ and $\widetilde{\boldsymbol{D}_{\beta_i}} = \text{diag}(\boldsymbol{D}_{\beta_i}, \cdots, \boldsymbol{D}_{\beta_i})$. Using a standard result from multivariate statistics (Anderson, 1958), we have that $f(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i | \boldsymbol{y}_i, \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i)$ is therefore also multivariate normal (see Supplementary Material for specification of the mean and covariance), which is simple to sample from. For $f(\boldsymbol{\gamma}_i | \boldsymbol{y}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\tau}_i)$ and $f(\tau_i | \boldsymbol{y}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i)$, we use (6) and (7) respectively.

### *2.3. MCEM Algorithm Summary*

Having derived a process for calculating the required maximum likelihood estimators and conditional expectations, we are now in a position to summarise the complete MCEM algorithm for estimating the model parameters for the skew-$t$-normal multi-level reduced-rank FPCA model.

A procedure for initialisation is described in the Supplementary Material. After initialisation, the algorithm alternates between approximating the conditional expectations by running the Gibbs sampler for each variable at the E-step, and calculating the maximum likelihood estimators with their sufficient statistics replaced by the conditional expectations. Note that the maximum likelihood estimators of $\boldsymbol{\theta}_\mu$, and the individual columns of $\boldsymbol{\Theta}_\alpha$ and $\boldsymbol{\Theta}_{\beta_i}$ are interdependent. Therefore it is necessary to employ an Expectation *Conditional* Maximisation (ECM) algorithm where each part of the M-step is carried out holding all of the other parameters fixed. In other words, first $\boldsymbol{\theta}_\mu$ is estimated holding $\boldsymbol{\Theta}_\alpha$ and $\boldsymbol{\Theta}_{\beta_i}$ fixed. Next, the first column of $\boldsymbol{\Theta}_\alpha$ is estimated, holding all other columns of $\boldsymbol{\Theta}_\alpha$ fixed along with $\boldsymbol{\Theta}_{\beta_i}$ and $\boldsymbol{\theta}_\mu$, and so on. This process can be iterated as in James, Hastie and Sugar (2000) or performed once per M-step as in Zhou et al. (2010). We have found the latter to work better in practice, in terms of numerical stability.

The maximum likelihood estimates of $\boldsymbol{\Theta}_\alpha$ and $\boldsymbol{\Theta}_{\beta_i}$ are not guaranteed to satisfy the constraint that the columns are orthogonal. While it is trivial to orthogonalise them by carrying out an eigen-decomposition, there is a lack of consensus as to whether this should be carried out at every iteration as in Zhou, Huang and Carroll (2008); Zhou et al. (2010) or once the EM algorithm has converged as in James, Hastie and Sugar (2000). In the case of the former we sacrifice the monotonicity property of the EM algorithm in order to ensure that the estimates remain within the valid parameter space at each iteration. In the case of the latter the monotonicity property is retained but intuition suggests that the algorithm may converge to parameter values that are far from those which maximise the constrained likelihood once the orthogonalisation is carried out. In fact, in our experience with the model under the assumption of normality, orthogonalising at each iteration results in the algorithm converging to a marginally larger likelihood at the expense of many more iterations required before convergence, possibly due to some of the iterations decreasing the likelihood. Furthermore, we have found that when the full-rank model is fit, such that $K$ and $L_i$ are set to the maximum permitted by the choice of spline basis,

computational instability can occur when orthogonalising at each iteration. An alternative to either of these approaches is to carry out the maximisation within the constrained parameter space - i.e. all matrices with orthonormal columns - as in Peng and Paul (2009). Such methods are difficult to implement due to their mathematical complexity and the success of James, Hastie and Sugar (2000); Zhou, Huang and Carroll (2008); Zhou et al. (2010) suggest that, practically speaking, they are unnecessary.

The complete algorithm is as follows:

1. Initialise parameters as described in the Supplementary Material
2. E-step: For each variable, run the Gibbs sampler for $S$ iterations, sampling from $f(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i | \boldsymbol{y}_i, \boldsymbol{\tau}_i, \boldsymbol{\gamma}_i)$, $f(\boldsymbol{\tau}_i | \boldsymbol{y}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i)$ and $f(\boldsymbol{\gamma}_i | \boldsymbol{y}_i, \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\tau}_i)$ in turn
3. M-step Step 1: Find the maximum likelihood estimators of the parameters of the skew-$t$-normal distributions using a simplex optimisation
4. M-step Step 2: Update $\widehat{\boldsymbol{D}_{\beta_i}}$, $i = 1, \cdots, M$
5. M-step Step 3: Update $\widehat{\sigma_i^2}$, $i = 1, \cdots, M$
6. M-step Constrained Maximisation Step 1: Update $\widehat{\boldsymbol{\theta}_\mu}$ while holding $\widehat{\boldsymbol{\Theta}_\alpha}$ and $\widehat{\boldsymbol{\Theta}_{\beta_i}}$, $i = 1, \cdots, M$ fixed
7. M-step Constrained Maximisation Step 2: Update $\widehat{\boldsymbol{\Theta}_\alpha}$ while holding $\widehat{\boldsymbol{\theta}_\mu}$ and $\widehat{\boldsymbol{\Theta}_{\beta_i}}$, $i = 1, \cdots, M$ fixed
8. M-step Constrained Maximisation Step 3: Update $\widehat{\boldsymbol{\Theta}_{\beta_i}}$, $i = 1, \cdots, M$ while holding $\widehat{\boldsymbol{\theta}_\mu}$ and $\widehat{\boldsymbol{\Theta}_\alpha}$ fixed
9. Check for convergence. If not converged, return to 3.
10. Orthogonalise $\widehat{\boldsymbol{\Theta}_\alpha}$ and $\widehat{\boldsymbol{\Theta}_{\beta_i}}$, $i = 1, \cdots, M$

### *2.4. Model Selection*

Two remaining issues are how to select the number of principal components, both at the variable- and replicate-level, and what spline basis to use. For selecting the number of principal components, two main approaches can be considered. In the first, the proportion of variance explained by each principal component function can be approximated by fitting the full-rank model - such that $K$ and $L_i$ (for all $i$) are the maximum permitted by the number of design time points (James, Hastie and Sugar, 2000). For the second method, cross-validation is used to score each potential value of $K$ and $L_i$ (Zhou et al., 2010). Note that both of these approaches require a subjective interpretation. For the proportion of variance explained method, either an arbitrary cutoff for cumulative variance such as 95% or 99% must be used, perhaps aided by an examination of a scree plot and a visualisation of the principal component functions in order to ascertain their interpretability. On the other hand, as Zhou et al. (2010) discuss, when using the cross-validation method on real data the score may simply decrease as the number of principal components increases, which results in always selecting the full-rank model going by the score alone. Therefore they suggest to instead subjectively trade-off between parsimony and cross-validation score, again with the aid of a scree plot. However, in the model of Zhou et al. (2010),

the number of principal components at the second-level is not dependent on the first-level and so the cross-validation method is much more tractable in their setting than ours, where the high-dimensional optimisation renders the approach impractical. By necessity therefore, we employ the proportion of variance explained method. On simulated data we have discovered that this works very well at identifying the correct number of principal components at the variable-level, but struggles at the replicate-level, most likely due to the much smaller sample sizes. We therefore suggest that a (much) more conservative criteria be applied at the replicate-level. For instance, we have found that retaining those principal components that explain 99% of the variance at the variable-level and 60% of the variance at the replicate-level worked well on these data sets.

For selecting the spline basis we suggest to use natural cubic splines with a knot placed at each design time point for several reasons. Firstly, this avoids the computational burden of having to select the number of knots. Secondly, many of the data sets provided by our collaborators have unequally spaced time points, and it therefore makes sense to use each time point as a knot in order to adequately capture the temporal dynamics. Thirdly, the reason to countenance against such an approach would be the danger of overfitting. This is accounted for through the use of the reduced-rank model, essentially placing a rank-constraint on the covariance matrix of the spline basis coefficients. Conceptually speaking, this approach to spline basis selection is quite similar to the use of smoothing splines, where a knot is placed at each design time point and overfitting is avoided through the use of a penalty parameter on the likelihood.

## 3. Results

### 3.1. Simulation comparison of the Gaussian single- and multi-level reduced-rank FPCA models

We set out to determine whether single-level approaches that assume the variables are independent are adequate when it comes to estimating the true underlying curves, and to quantify the improvement that can be gained through the use of a multi-level model using the following simulation setting.

We generated data under the multi-level reduced-rank FPCA model (2) with normality assumed, as the skew-$t$-normal model is too computationally intensive, with its reliance on MC methods, for large scale simulation studies. We fixed the number of principal components at the variable level to $K = 2$ with a single principal component at the replicate level for each variable, so that $L_i = 1$ for all $i$. We used a B-spline basis with a single knot placed at the centre of the time course. The spline coefficients for the grand mean, $\boldsymbol{\theta}_\mu$, were fixed to produce the curve that can be seen in Figure 3.

The spline coefficients for the variable-level principal component functions, $\boldsymbol{\theta}_{\alpha_1}$ and $\boldsymbol{\theta}_{\alpha_2}$ were chosen in order to produce the simulated curves given in Figure 2. Visualising the grand mean plus and minus each principal component function is a typical way of helping to understand their effect (Ramsay and

Silverman, 2005). These plots are given in Figure 3. The solid line is the grand mean. The points denoted by '+' are the function $\mu(t) + C\zeta_k(t)$ evaluated on a coarse grid of points where $C$ is some constant responsible for scaling the principal component function and subjectively chosen in order to aid clarity of the visualisation. Similarly, the points denoted by '$-$' are the same except for $\mu(t) - C\zeta_k(t)$. From these plots it should be clear that the first principal component has the effect of rotating the first half of the time course, which alters both the level at which the time course begins and the level to which it peaks. To a lesser extent the exact time at which the peak occurs is also affected. On the other hand, the second principal component function rotates the second half of the time course, thereby controlling whether the curve levels off, continues to decrease or starts to increase after the peak and the dip.

For the replicate-level principal components, the same principal component function was fixed for all variables. Note that this only serves to simplify the simulation scheme, and the multi-level reduced-rank FPCA model will still estimate the between-replicate variation for each variable independently. By choosing the spline coefficients $\boldsymbol{\theta}_{\beta_1}$ to produce the profile given on the left hand side of Figure 4, the effect of the principal component function is to scale the height to which the curves peak, as shown on the right hand side of the same Figure.

The remaining parameters to be determined are the variance components $\boldsymbol{D}_\alpha$, $\boldsymbol{D}_{\beta_i}$ and $\sigma_i^2$ for all $i$. We set $\boldsymbol{D}_\alpha = \text{diag}(0.3, 0.1)$ so that 75% of the variance at the variable level is explained by the first principal component function. The single element of $\boldsymbol{D}_{\beta_i}$ was set to 0.075 for all $i$ so that the level of between-replicate variance is less than that of the between-variable. Similarly, the noise was set to $\sigma_i^2 = 0.05$ for all $i$. As before, this only serves to simplify the simulation scheme and separate estimates will be made for $\boldsymbol{D}_{\beta_i}$ and $\sigma_i^2$ for each variable. The simulation scheme described here is flexible enough to produce a wide range of believable curves as evidenced by the examples given in Supplementary Figure 5.

In order to compare the single- and multi-level reduced-rank FPCA models under a range of experimental designs and to determine whether standard practice is appropriate, we generated different data sets by varying the number of replicates as either 5, 10 or 20. Roughly speaking, with regards real data sets, these correspond to realistic, less frequent and unrealistic numbers of replicates respectively. The number of time points per data set was fixed to 5. We focused on evaluating the ability to estimate the variable-level curves and for this reason we also varied the number of variables between 100, $1,000$ and $10,000$. Although broadly speaking $10,000$ is the only realistic value of the three for 'omics data sets (unless the variables have been subject to some initial filtering procedure), we were interested in exploring the properties of the multi-level model and determining how many variables are required to adequately assess the between-variable variation.

We generated $1,000$ data sets under each condition. For each data set we fit both the single- and multi-level reduced-rank FPCA models under the assumption of normality. For the purposes of this study we chose not to consider the issue of correctly selecting the number of principal component functions at

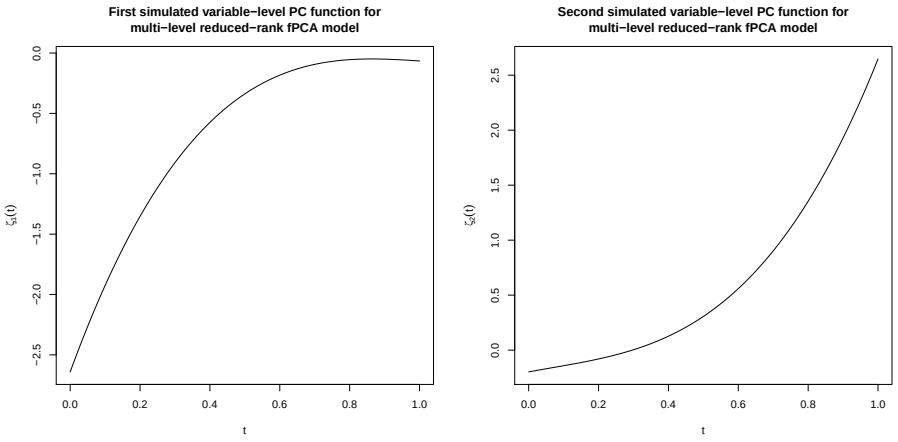



FIG 2. *Two simulated variable-level principal component functions used in our simulation study to compare the Gaussian single- and multi-level reduced-rank FPCA models. The principal component functions were set to explain* 75% *and* 25% *of the variance in the data respectively.*

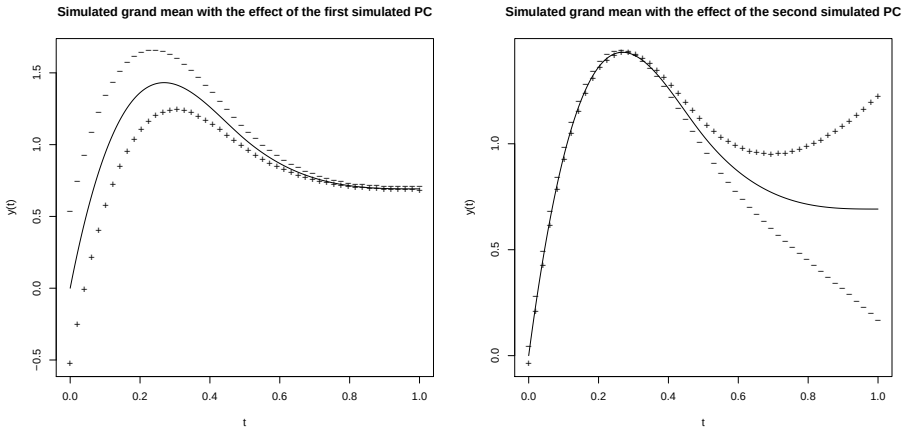



FIG 3. *The effect of the two simulated principal component functions on the simulated grand mean from our simulation study to compare the Gaussian single- and multi-level reduced-rank FPCA models. The first principal component function rotates the first half of the time course, therefore affecting the height to which the functions peak, the level at which the time course begins and, to a lesser extent, the exact time at which the peak occurs. Conversely the second principal component function rotates the second half of the time course, thereby controlling whether the curve levels off by the end of the time course, continues to decrease or starts to increase.*

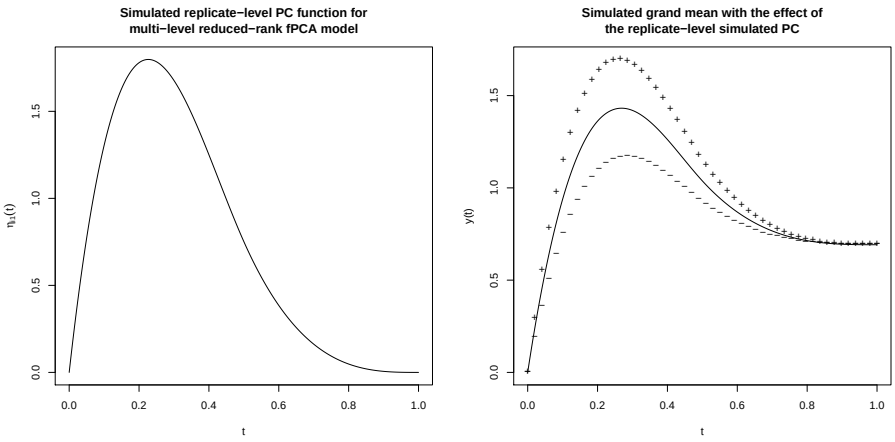



FIG 4. *Profile of the simulated replicate-level principal component function used in our simulation study to compare the Gaussian single- and multi-level reduced-rank FPCA models, left, and its effect on the grand mean, right. The principal component function has the effect of scaling the height to which the curve peaks.*

the variable- and replicate-levels and the number or location of the knots of the spline basis and simply input the correct values to each algorithm. We compared each variable-level curve by discretising it on a fine grid of points and calculating the mean squared error between it and the underlying true curve. We then averaged this error across all variables and all data sets to give a single measure for each model for each condition.

The complete results of the simulation study are presented in Supplementary Tables 3 and 4. In all scenarios the multi-level model substantially improves upon the single-level model. For the multi-level model, doubling the number of replicates roughly halves the estimation error. However, increasing the number of variables has a much less pronounced effect, suggesting the multi-level model still performs well when data sets have been pre-filtered. For the single-level model, doubling the number of replicates similarly roughly halves the estimation error. As expected, increasing the number of variables for this model has no effect on the estimation error as each variable is fit in isolation. The most salient insight to glean from these results is in the comparison between the two tables. Comparing first the case of 5 replicates and $10,000$ variables for the two models, the multi-level model offers an improvement in estimation error of approximately a factor of ten. Considering the single-level model alone, in order to attain an equivalent improvement in estimation error, the number of replicates needs to quadruple from 5 to 20. These results suggest that simply using a more sophisticated model has the potential to dramatically reduce experimental costs.

### 3.2. Real data analysis

We fit the skew-$t$-normal multi-level reduced-rank FPCA model to a genomics data set study the genetic response to infection by BCG, the vaccine for tuberculosis, in 9 human volunteers. We determined that the MCEM algorithm had converged after 1050 iterations by examining parameter trace plots. On the right hand side of Figure 1 we show the fit obtained under the new model to the CCL20 transcript. As can be clearly seen, the mean curve is much more reasonable, closely following the underlying observations.

The variable-level principal component functions are given in Figure 5. The first principal component – which explains a huge proportion of the variance, 99.998% – is responsible for vertically shifting a given transcript. In this respect, it is the most uninteresting of the principal component functions as it does not control the shape of the curves. We stress that it is not suprising that such a large proportion of the variance is explained by vertical shifts given that very few of the tens of thousands of variables in an 'omics data set will actually exhibit significant changes over time. It is therefore still instructive to consider the other principal component functions even if the proportion of variance they explain may lead to them being overlooked in more traditional PCA application areas.

The second principal component function accounts for those transcripts which rapidly spike before plateauing at some elevated level of expression or, conversely, are rapidly repressed. The CCL20 transcript is a example of this. The third principal component function describes those transcripts which are induced or repressed more slowly, tailing off at around 7 or 8 hours. In Supplementary Figure 6 we give an example of a transcript that exhibits this profile and was found to have the highest positive loading on the third principal component function. More complex profiles can be explained by a combination of these two components. The fourth principal component function is less interpretable but appears to be mainly controlling for variation at the end of the time course. The fifth principal component function is even less interpretable and should probably be discarded when the data is fit for a second time under the reduced-rank model.

## 4. Discussion

To date, multi-level functional models such as the one we have presented here have yet to be applied in the 'omics fields that we focus on. In other domains, Zhou et al. (2010) independently developed a similar multi-level reduced-rank FPCA model under the normality assumption. Their model differs from ours in a few key respects. Firstly, the second-level prinicpal component loadings (in our model, this would be the replicate-level) are not specific to the first-level. In other words, in our context, their model would assume that the within-variable variation is identical for all variables. As we motivated with Supplementary Figures 1 and 2, real data does not support this assumption. Secondly, they

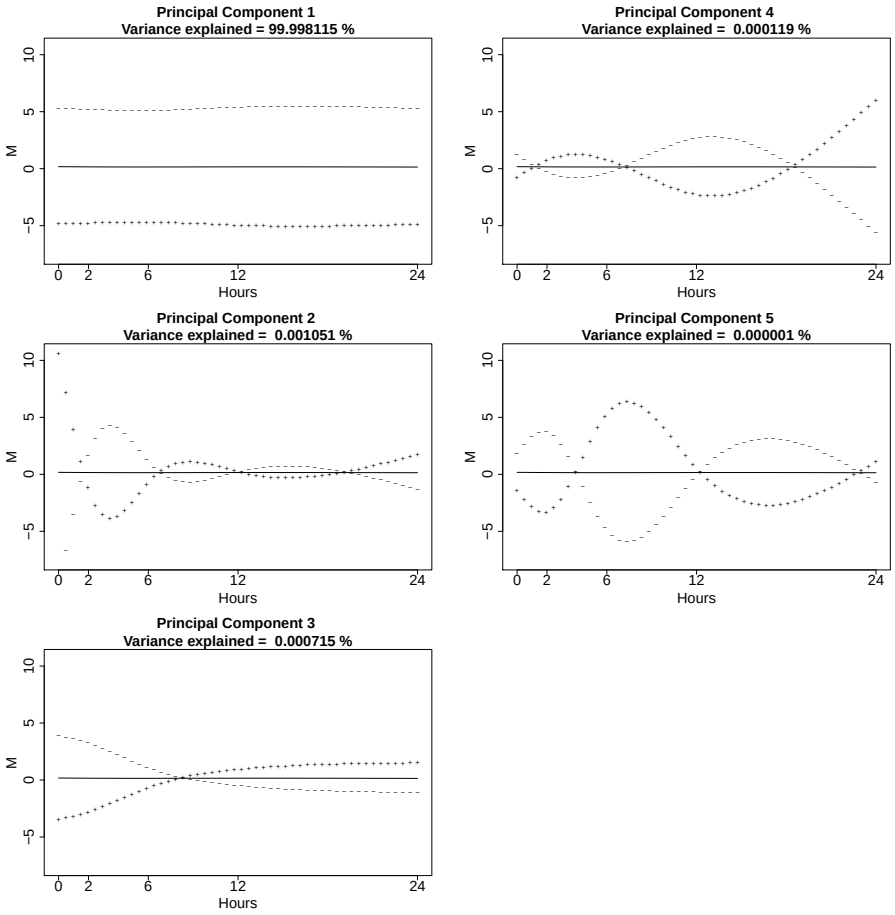



FIG 5. *Variable-level principal component functions obtained from fitting the skew-t-normal multi-level reduced-rank fPCA model to our example real data set. The solid line is the grand mean across all transcripts. The points '+' indicate the effect of adding each principal component function multiplied by a constant C to the grand mean, where C is simply subjectively chosen to aid the visualisation. Similarly the points '-' indicate the effect of subtracting each principal component function multiplied by C from the grand mean. See discussion in the main text.*

allow for the replicate-level loadings to be correlated within a given variable. In this respect their model is an extension of ours; however, this is motivated by spatial dependencies in their case study that do not exist in our example data sets. Thirdly, they use a penalised spline representation for the principal component functions. In principle this should allow for a more data-driven approach to smoothing than our choice of natural splines with the maximum number of knots. However, we note that their data set has many more 'time' points (in fact, the dependent variable is distance), than our case studies ($20-40$) and therefore smoothing is more likely to be an issue if the function has been oversampled. Furthermore, they select only three distinct smoothing parameters, one for the grand mean, one for all variable-level principal component functions, and one for all replicate-level principal component functions. Although it is clear to understand the computational burden that motivates such a restriction, it seems to do away with the advantage that motivates penalised estimation in the first place, which is to account for principal component functions of varying smoothness. Fourthly and finally, they fit a single error variance for all variables. It is well-known that noise in microarray experiments is transcript-dependent (Tusher, Tibshirani and Chu, 2001) and so such a noise model would be inadequate for our purposes. Finally, the model is demonstrated on a data set with far fewer variables than our case studies (3 comapred with on the order of $10,000$) which may explain why they did not experience the same problems with assuming normality that we did.

Although we have demonstrated that the MCEM algroithm for the skew-$t$-normal model yields biologically interpretable results on real data, the computational burden is severe. We suggest several lines of attack for developing a practical algorithm that can fit the model in a more reasonable time frame. Firstly, a different sampling scheme could be employed in the MCEM algorithm, specifically importance sampling using a multivariate $t$ proposal with a small degrees of freedom parameter. Moving away from the EM-algorithm, there may be the potential for an (approximate) analytical solution. In particular, the work of Forchini (2008) on deriving the form of the density of the sum of a normal and a Student-$t$ distributed random variable may provide some guidance. Deriving the density of the sum of a normal and a skew-$t$-normal random variable would be an important first step. However, the result of Forchini (2008) relies on truncating a Taylor series expansion and so careful consideration should be given as to the accuracy of such an approximation. Ultimately, however, we believe that the most fruitful line of further investigation is a fully Bayesian approach, taken by imposing prior distributions on the model parameters. Parameter estimation could then be carried out using a modified version of our Gibbs sampler. Jara, Quintana and Martin (2008) have already demonstrated that such a Bayesian approach works well for single-level mixed-effects models with either skew-$t$ or skew-normal random-effects.

## References

Anderson, T. W. (1958). *Introduction to Multivariate Statistical Analysis.*

Wiley.

Angelini, C., Canditiis, D. D. and Pensky, M. (2009). Bayesian models for two-sample time-course microarray experiments. *Computational Statistics & Data Analysis* **53** 1547 - 1565. Statistical Genetics & Statistical Genomics: Where Biology, Epistemology, Statistics, and Computation Collide.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society, Series B* **65** 367-389.

Berk, M., Ebbels, T. and Montana, G. (2011). A statistical framework for metabolic profiling using longitudinal data. *Bioinformatics* **27** 1979 - 1985.

Berk, M., Montana, G., Levin, M. and Hemingway, C. (2010). Longitudinal analysis of gene expression profiles using functional mixed-effects models. In *Studies in Theoretical and Applied Statistics*.

Di, C., Crainiceanu, C. M., Kuechenhoff, H. and Peters, A. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics* **3** 458 - 488.

Forchini, G. (2008). The distribution of the sum of a normal and a t random variable with arbitrary degrees of freedom. *METRON - International Journal of Statistics* **2** 205-208.

Gómez, H. W., Venegas, O. and Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics* **18** 395–407.

Ho, H. J. and Lin, T.-I. (2010). Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal* **52** 449 - 469.

James, G., Hastie, T. and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587-602.

Jara, A., Quintana, F. and Martin, E. S. (2008). Linear mixed models with skew-elliptical distributions: A Bayesian approach. *Computational Statistics & Data Analysis* **52** 5033 - 5045.

Liu, X. and Yang, M. C. K. (2009). Identifying temporally differentially expressed genes through functional principal components analysis. *Biostat* kxp022.

Ma, P., Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Res* **34** 1261–1269.

Montana, G., Berk, M. and Ebbels, T. (2011). *Software Tools and Algorithms for Biological Systems* Modelling short time series in metabolomics: a functional data analysis approach 307-316. Springer, New York.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal* **7** 308-313.

Peng, J. and Paul, D. (2009). A Geometric Approach to Maximum Likelihood Estimation of the Functional Principal Components From Sparse Longitudinal Data. *Journal of Computational and Graphical Statistics* **18** 995-1015.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*, 2 ed. Springer, New York.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A* **102** 12837–12842.

Tai, Y. C. and Speed, T. P. (2009). On Gene Ranking Using Replicated Microarray Time Course Data. *Biometrics* **65** 40-51.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98** 5116-5121.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association* **85** pp. 699-704.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association* **100** 577-590.

Zhou, L., Huang, J. Z. and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95** 601-619.

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V. and Carroll, R. J. (2010). Reduced Rank Mixed Effects Models for Spatially Correlated Hierarchical Functional Data. *Journal of the American Statistical Association* **105** 390-400.

# Supplementary Material for "A Skew-$t$-Normal Multi-Level Reduced-Rank Functional PCA Model with Applications to Replicated 'Omics Time Series Data Sets"

## 1 The Gaussian multi-level reduced-rank FPCA model

For the Gaussian case we take the following distributional assumptions:

$$\boldsymbol{\alpha}_i \overset{i.i.d.}{\sim} MVN(\mathbf{0}, \boldsymbol{D}_\alpha) \quad \boldsymbol{\beta}_{ij} \overset{i.i.d.}{\sim} MVN(\mathbf{0}, \boldsymbol{D}_{\beta_i}) \quad \boldsymbol{\epsilon}_{ij} \overset{i.i.d.}{\sim} MVN(\mathbf{0}, \sigma_i^2 \boldsymbol{I}_{N_{ij} \times N_{ij}}) \tag{1}$$

where the $K \times K$ and $L_i \times L_i$ matrices $\boldsymbol{D}_\alpha$ and $\boldsymbol{D}_{\beta_i}$ are both assumed to be diagonal otherwise they would be confounded with $\boldsymbol{\Theta}_\alpha$ and $\boldsymbol{\Theta}_{\beta_i}$. Furthermore, we assume that $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are all independent of each other for all $i$ and $j$. To enforce the orthogonality constraint on the functions $\zeta_k(t)$ and $\eta_{il}(t)$, in addition to transforming the spline basis, we also impose that $\boldsymbol{\Theta}_\alpha^T \boldsymbol{\Theta}_\alpha = \boldsymbol{I}_{K \times K}$ and $\boldsymbol{\Theta}_{\beta_i}^T \boldsymbol{\Theta}_{\beta_i} = \boldsymbol{I}_{L_i \times L_i}$.

Under the distributional assumptions given above, $\boldsymbol{y}_i$ is marginally distributed as

$$\boldsymbol{y}_i \sim MVN(\boldsymbol{B}_i \boldsymbol{\theta}_\mu, \boldsymbol{V}_i)$$

where

$$\begin{aligned} \boldsymbol{V}_i &= \boldsymbol{B}_i \boldsymbol{\Theta}_\alpha \boldsymbol{D}_\alpha \boldsymbol{\Theta}_\alpha^T \boldsymbol{B}_i^T + \widetilde{\boldsymbol{B}_i} \widetilde{\boldsymbol{\Theta}_{\beta_i}} \widetilde{\boldsymbol{D}_{\beta_i}} \widetilde{\boldsymbol{\Theta}_{\beta_i}^T} \widetilde{\boldsymbol{B}_i^T} + \sigma_i^2 \boldsymbol{I}_{N_i \times N_i} \\ \widetilde{\boldsymbol{D}_{\beta_i}} &= \mathrm{diag}(\boldsymbol{D}_{\beta_i}, \cdots, \boldsymbol{D}_{\beta_i}) \end{aligned} \tag{2}$$

As in James et al. (2000), the model parameters

$$\boldsymbol{\psi}_i = \left\{ \boldsymbol{\theta}_\mu, \boldsymbol{\Theta}_\alpha, \boldsymbol{\Theta}_{\beta_i}, \boldsymbol{D}_\alpha, \boldsymbol{D}_{\beta_i} \sigma_i^2 \right\} \quad i = 1, \cdots, M$$

where $M$ is the total number of variables in the data set, can be estimated by treating the principal component loadings as missing data and employing the EM algorithm. Under the distributional assumptions given above, the complete data log-likelihood is

$$\begin{aligned} \sum_{i=1}^{M} \mathcal{L}(\boldsymbol{\psi}_i | \boldsymbol{y}_i) &= \sum_{i=1}^{M} \left[ \log f(\boldsymbol{y}_i | \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\psi}_i) + \log f(\boldsymbol{\alpha}_i | \boldsymbol{\psi}_i) + \sum_{j=1}^{n_i} \log f(\boldsymbol{\beta}_{ij} | \boldsymbol{\psi}_i) \right] \\ &= C - \frac{1}{2} \sum_{i=1}^{M} \Bigg[ N_i \log \sigma_i^2 + \sigma_i^{-2} || \boldsymbol{y}_i - \boldsymbol{B}_i \boldsymbol{\theta}_\mu - \boldsymbol{B}_i \boldsymbol{\Theta}_\alpha \boldsymbol{\alpha}_i - \widetilde{\boldsymbol{B}_i} \widetilde{\boldsymbol{\Theta}_{\beta_i}} \boldsymbol{\beta}_i ||^2 + \\ &\qquad \log |\boldsymbol{D}_\alpha| + \boldsymbol{\alpha}_i^T \boldsymbol{D}_\alpha^{-1} \boldsymbol{\alpha}_i + \sum_{j=1}^{n_i} \left[ \log |\boldsymbol{D}_{\beta_i}| + \boldsymbol{\beta}_{ij}^T \boldsymbol{D}_{\beta_i}^{-1} \boldsymbol{\beta}_{ij} \right] \Bigg] \end{aligned} \tag{3}$$

where $C$ is an additive constant. We will now proceed to derive the maximum likelihood estimators of the model parameters, followed by the required conditional expectations for the EM algorithm.

## 1.1 MLE of $\boldsymbol{\theta}_\mu$

To derive the maximum likelihood estimator of $\boldsymbol{\theta}_\mu$ first ignore all irrelevant terms in (3) and for succinctness write $\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu - \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\boldsymbol{\beta}_i = \boldsymbol{y}_i^* - \boldsymbol{B}_i\boldsymbol{\theta}_\mu$. Then the partial derivative with respect to $\boldsymbol{\theta}_\mu$ yields

$$\frac{\partial -\frac{1}{2}\sum_{i=1}^{M}\sigma_i^{-2}(\boldsymbol{y}_i^* - \boldsymbol{B}_i\boldsymbol{\theta}_\mu)^T(\boldsymbol{y}_i^* - \boldsymbol{B}_i\boldsymbol{\theta}_\mu)}{\partial\boldsymbol{\theta}_\mu} = \sum_{i=1}^{M}\sigma_i^{-2}\left[-2\boldsymbol{B}_i^T\boldsymbol{y}_i^* + 2\boldsymbol{B}_i^T\boldsymbol{B}_i\boldsymbol{\theta}_\mu\right]$$

Equating to zero and solving for $\boldsymbol{\theta}_\mu$ gives

$$\begin{aligned}
\sum_{i=1}^{M}(\boldsymbol{B}_i^T\boldsymbol{B}_i)\boldsymbol{\theta}_\mu &= \sum_{i=1}^{M}\boldsymbol{B}_i^T\boldsymbol{y}_i^* \\
\widehat{\boldsymbol{\theta}_\mu} &= (\sum_{i=1}^{M}\boldsymbol{B}_i^T\boldsymbol{B}_i)^{-1}\sum_{i=1}^{M}\boldsymbol{B}_i^T\boldsymbol{y}_i^* \\
\widehat{\boldsymbol{\theta}_\mu} &= (\sum_{i=1}^{M}\boldsymbol{B}_i^T\boldsymbol{B}_i)^{-1}\sum_{i=1}^{M}\boldsymbol{B}_i^T\left[\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\boldsymbol{\beta}_i\right] \qquad (4)
\end{aligned}$$

## 1.2 MLE of $\boldsymbol{\theta}_{\alpha_k}$

For the maximum likelihood estimator of $\boldsymbol{\theta}_{\alpha_k}$, which recall is the $k$-th column of $\boldsymbol{\Theta}_\alpha$, proceed in exactly the same way as for $\boldsymbol{\theta}_\mu$. For clarity first redefine $\boldsymbol{y}_i^* = \boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu - \boldsymbol{B}_i\sum_{k'\neq k}\left[\boldsymbol{\theta}_{\alpha_{k'}}\alpha_{ik'}\right] - \widetilde{\boldsymbol{\Theta}_{\beta_i}}\boldsymbol{\beta}_i$. Then, taking the partial derivative of the relevant terms of (3) with respect to $\boldsymbol{\theta}_{\alpha_k}$ gives

$$\begin{aligned}
&\frac{\partial -\frac{1}{2}\sum_{i=1}^{M}\sigma_i^{-2}(\boldsymbol{y}_i^* - \boldsymbol{B}_i\boldsymbol{\theta}_{\alpha_k}\alpha_{ik})^T(\boldsymbol{y}_i^* - \boldsymbol{B}_i\boldsymbol{\theta}_{\alpha_k}\alpha_{ik})}{\partial\boldsymbol{\theta}_{\alpha_k}} = \\
&\qquad\qquad \sum_{i=1}^{M}\sigma_i^{-2}\Big[-2\alpha_{ik}\boldsymbol{B}_i^T\boldsymbol{y}_i^* + 2\alpha_{ik}^2\boldsymbol{B}_i^T\boldsymbol{B}_i\boldsymbol{\theta}_\mu\Big]
\end{aligned}$$

Equating to zero and solving for $\boldsymbol{\theta}_{\alpha_k}$ gives

$$\begin{aligned}
\sum_{i=1}^{M}(\alpha_{ik}^2\boldsymbol{B}_i^T\boldsymbol{B}_i)\boldsymbol{\theta}_{\alpha_k} &= \sum_{i=1}^{M}\alpha_{ik}\boldsymbol{B}_i^T\boldsymbol{y}_i^* \\
\widehat{\boldsymbol{\theta}_{\alpha_k}} &= (\sum_{i=1}^{M}\alpha_{ik}^2\boldsymbol{B}_i^T\boldsymbol{B}_i)^{-1}\sum_{i=1}^{M}\alpha_{ik}\boldsymbol{B}_i^T\boldsymbol{y}_i^*
\end{aligned}$$

and so

$$\widehat{\boldsymbol{\theta}_{\alpha_k}} = (\sum_{i=1}^{M}\alpha_{ik}^2\boldsymbol{B}_i^T\boldsymbol{B}_i)^{-1}\sum_{i=1}^{M}\alpha_{ik}\boldsymbol{B}_i^T\left[\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu - \boldsymbol{B}_i\sum_{k'\neq k}\left[\boldsymbol{\theta}_{\alpha_k}\alpha_{ik'}\right] - \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\boldsymbol{\beta}_i\right] \qquad (5)$$

## 1.3 MLE of $\boldsymbol{\theta}_{\beta_{il}}$

The derivation for the maximum likelihood estimator of $\boldsymbol{\theta}_{\beta_{il}}$ is similar to that for $\boldsymbol{\theta}_{\alpha_k}$ except this time we only need to consider observations on variable $i$. Hence we define $\boldsymbol{y}_{ij}^* = \boldsymbol{y}_{ij} - \boldsymbol{B}_{ij}\boldsymbol{\theta}_\mu - \boldsymbol{B}_{ij}\boldsymbol{\Theta}_{\alpha_i} - \boldsymbol{B}_{ij}\sum_{l'\neq l}\left[\boldsymbol{\theta}_{\beta_{il'}}\beta_{ijl'}\right]$

and take the partial derivative of the relevant terms in (3) with respect to $\boldsymbol{\theta}_{\beta_{il}}$ which yields

$$\frac{\partial -\frac{1}{2}\sum_{j=1}^{n_i}\sigma_i^{-2}(\boldsymbol{y}_{ij}^* - \boldsymbol{B}_{ij}\boldsymbol{\theta}_{\beta_{il}}\beta_{ijl})^T(\boldsymbol{y}_i^* - \boldsymbol{B}_i\boldsymbol{\theta}_{\beta_{il}}\beta_{ijl})}{\partial \boldsymbol{\theta}_{\beta_{il}}} = \sum_{j=1}^{n_i}\sigma_i^{-2}\Big[-2\beta_{ijl}\boldsymbol{B}_{ij}^T\boldsymbol{y}_{ij}^* + 2\beta_{ijl}^2\boldsymbol{B}_{ij}^T\boldsymbol{B}_{ij}\boldsymbol{\theta}_{\beta_{il}}\Big]$$

Equating to zero and solving for $\boldsymbol{\theta}_{\beta_{il}}$ gives

$$\sum_{j=1}^{n_i}(\beta_{ijl}^2\boldsymbol{B}_{ij}^T\boldsymbol{B}_{ij})\boldsymbol{\theta}_{\beta_{il}} = \sum_{j=1}^{n_i}\beta_{ijl}\boldsymbol{B}_{ij}^T\boldsymbol{y}_{ij}^*$$

$$\widehat{\boldsymbol{\theta}_{\beta_{il}}} = (\sum_{j=1}^{n_i}\beta_{ijl}^2\boldsymbol{B}_{ij}^T\boldsymbol{B}_{ij})^{-1}\sum_{j=1}^{n_i}\beta_{ijl}\boldsymbol{B}_{ij}^T\boldsymbol{y}_{ij}^*$$

and so

$$\widehat{\boldsymbol{\theta}_{\beta_{il}}} = (\sum_{j=1}^{n_i}\beta_{ijl}^2\boldsymbol{B}_{ij}^T\boldsymbol{B}_{ij})^{-1}\sum_{j=1}^{n_i}\beta_{ijl}\boldsymbol{B}_{ij}^T\left[\boldsymbol{y}_{ij} - \boldsymbol{B}_{ij}\boldsymbol{\theta}_\mu - \boldsymbol{B}_{ij}\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i - \boldsymbol{B}_{ij}\sum_{l'\neq l}[\boldsymbol{\theta}_{\beta_{il'}}\beta_{ijl'}]\right] \tag{6}$$

## 1.4 MLE of $\boldsymbol{D}_\alpha$

As $\boldsymbol{D}_\alpha$ is diagonal we consider estimating each diagonal element separately. The relevant terms in (3) are $-\frac{1}{2}\sum_{i=1}^{M}\left[\log|\boldsymbol{D}_\alpha| + \boldsymbol{\alpha}_i^T\boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_i\right]$. As $\boldsymbol{D}_\alpha$ is diagonal we can write

$$-\frac{1}{2}\sum_{i=1}^{M}\left[\log|\boldsymbol{D}_\alpha| + \boldsymbol{\alpha}_i^T\boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_i\right] = -\frac{1}{2}\sum_{i=1}^{M}\left[\log\prod_{k'=1}^{K}[\boldsymbol{D}_\alpha]_{k'k'} + \sum_{k'=1}^{K}\frac{\alpha_{ik'}^2}{[\boldsymbol{D}_\alpha]_{k'k'}}\right]$$

where $[\boldsymbol{D}_\alpha]_{k'k'}$ denotes the $k'$-th diagonal element of $\boldsymbol{D}_\alpha$. Considering the case of estimating the $k$-th diagonal element of $\boldsymbol{D}_\alpha$ we first separate out the relevant terms

$$-\frac{1}{2}\sum_{i=1}^{M}\left[\log\prod\mathrm{diag}(\boldsymbol{D}_\alpha) + \sum_{k'=1}^{K}\frac{\alpha_{ik'}^2}{[\boldsymbol{D}_\alpha]_{k'k'}}\right] = -\frac{1}{2}\sum_{i=1}^{M}\left[\log[\boldsymbol{D}_\alpha]_{kk} + \frac{\alpha_{ik}^2}{[\boldsymbol{D}_\alpha]_{kk}} + \sum_{k'\neq k}\left[\log[\boldsymbol{D}_\alpha]_{k'k'} + \frac{\alpha_{ik'}^2}{[\boldsymbol{D}_\alpha]_{k'k'}}\right]\right]$$

Then we take the partial derivative with respect to $[\boldsymbol{D}_\alpha]_{kk}$

$$\frac{\partial -\frac{1}{2}\sum_{i=1}^{M}\left[\log[\boldsymbol{D}_\alpha]_{kk} + \frac{\alpha_{ik}^2}{[\boldsymbol{D}_\alpha]_{kk}}\right]}{\partial[\boldsymbol{D}_\alpha]_{kk}} = -\frac{1}{2}\sum_{i=1}^{M}\left[\frac{1}{[\boldsymbol{D}_\alpha]_{kk}} - \frac{\alpha_{ik}^2}{[\boldsymbol{D}_\alpha]_{kk}^2}\right]$$

Equating to zero and solving for $[\boldsymbol{D}_\alpha]_{kk}$ gives

$$
\begin{aligned}
-\frac{1}{2}\sum_{i=1}^{M}\left[\frac{1}{[\boldsymbol{D}_\alpha]_{kk}}\right] &= -\frac{1}{2}\sum_{i=1}^{M}\left[\frac{\alpha_{ik}^2}{[\boldsymbol{D}_\alpha]_{kk}^2}\right] \\
M[\boldsymbol{D}_\alpha]_{kk} &= \sum_{i=1}^{M}\alpha_{ik}^2 \\
\widehat{[\boldsymbol{D}_\alpha]_{kk}} &= \frac{1}{M}\sum_{i=1}^{M}\alpha_{ik}^2
\end{aligned} \tag{7}
$$

## 1.5 MLE of $\boldsymbol{D}_{\beta_i}$

Following exactly the same procedure to dervive the maximum likelihood estimator of $\boldsymbol{D}_{\beta_i}$ gives

$$
\begin{aligned}
-\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{1}{[\boldsymbol{D}_{\beta_i}]_{ll}}\right] &= -\frac{1}{2}\sum_{j=1}^{n_i}\left[\frac{\beta ijl^2}{[\boldsymbol{D}_{\beta_i}]_{ll}^2}\right] \\
n_i[\boldsymbol{D}_{\beta_i}]_{ll} &= \sum_{j=1}^{n_i}\beta_{ijl}^2 \\
\widehat{[\boldsymbol{D}_{\beta_i}]_{ll}} &= \frac{1}{n_i}\sum_{j=1}^{n_i}\beta_{ijl}^2
\end{aligned} \tag{8}
$$

## 1.6 MLE of $\sigma_i^2$

To derive the maximum likelihood estimator of $\sigma_i^2$ first make the substitution $\boldsymbol{\epsilon}_i = \boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu - \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i - \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\boldsymbol{\beta}_i$ for clarity. Then the relevant terms of (3) are $-\frac{1}{2}\left[N_i\log\sigma_i^2 + \sigma_i^{-2}\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i\right]$. Taking the partial derivative with respect to $\sigma_i^2$ gives

$$
\frac{\partial -\frac{1}{2}\left[N_i\log\sigma_i^2 + \sigma_i^{-2}\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i\right]}{\partial\sigma_i^2} = -\frac{1}{2}\left[\frac{N_i}{\sigma_i^2} - \frac{1}{\sigma_i^4}\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i\right]
$$

Equating to zero and solving for $\sigma_i^2$ gives

$$
\widehat{\sigma_i^2} = \frac{1}{N_i}\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i \tag{9}
$$

## 1.7 Conditional Expectations

Supplementary Table 2 summarises the requried conditional expectations for the EM algorithm, based on the sufficient statistics of the maximum likelihood estimators derived above. In order to derive these conditional expectations we first write

$$
\begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \\ \boldsymbol{\epsilon}_i \\ \boldsymbol{y}_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{B}_i\boldsymbol{\theta}_\mu \end{bmatrix},
\begin{bmatrix}
\boldsymbol{D}_\alpha & \mathbf{0} & \mathbf{0} & \boldsymbol{D}_\alpha\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T \\
\mathbf{0} & \widetilde{\boldsymbol{D}_{\beta_i}} & \mathbf{0} & \widetilde{\boldsymbol{D}_{\beta_i}}\widetilde{\boldsymbol{\Theta}_{\beta_i}^T}\widetilde{\boldsymbol{B}_i^T} \\
\mathbf{0} & \mathbf{0} & \sigma_i^2\boldsymbol{I}_{N_i\times N_i} & \sigma_i^2\boldsymbol{I}_{N_i\times N_i} \\
\boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{D}_\alpha & \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widetilde{\boldsymbol{D}_{\beta_i}} & \sigma_i^2\boldsymbol{I}_{N_i\times N_i} & \boldsymbol{V}_i
\end{bmatrix}\right)
$$

| Conditional Expectation | Parameters required for |
|---|---|
| $E[\boldsymbol{\alpha}_i\|\boldsymbol{y}_i]$ | $\boldsymbol{\theta}_\mu, \sigma_i^2$ |
| $E[\boldsymbol{\beta}_i\|\boldsymbol{y}_i]$ | $\boldsymbol{\theta}_\mu, \sigma_i^2$ |
| $E[\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T\|\boldsymbol{y}_i]$ | $\boldsymbol{\Theta}_\alpha, \boldsymbol{D}_\alpha$ |
| $E[\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T\|\boldsymbol{y}_i]$ | $\boldsymbol{\Theta}_{\beta_i}, \boldsymbol{D}_{\beta_i}$ |
| $E[\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T\|\boldsymbol{y}_i]$ | $\boldsymbol{\Theta}_\alpha, \boldsymbol{\Theta}_{\beta_i}$ |
| $E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i\|\boldsymbol{y}_i]$ | $\sigma_i^2$ |

Supplementary Table 1: The required conditional expectations for the EM algorithm for the Gaussian multi-level reduced-rank fPCA model

| Conditional Expectation | Parameters required for |
|---|---|
| $E[\boldsymbol{\alpha}_i\|\boldsymbol{y}_i]$ | $\boldsymbol{\theta}_\mu, \sigma_i^2, \xi_{\alpha_k}, \sigma^2_{\alpha_k}, \lambda_{\alpha_k}, \nu_{\alpha_k}$ |
| $E[\boldsymbol{\beta}_i\|\boldsymbol{y}_i]$ | $\boldsymbol{\theta}_\mu, \sigma_i^2$ |
| $E[\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T\|\boldsymbol{y}_i]$ | $\boldsymbol{\Theta}_\alpha$ |
| $E[\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T\|\boldsymbol{y}_i]$ | $\boldsymbol{\Theta}_{\beta_i}, \boldsymbol{D}_{\beta_i}$ |
| $E[\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T\|\boldsymbol{y}_i]$ | $\boldsymbol{\Theta}_\alpha, \boldsymbol{\Theta}_{\beta_i}$ |
| $E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i\|\boldsymbol{y}_i]$ | $\sigma_i^2$ |

Supplementary Table 2: The required conditional expectations for the EM algorithm for the skew-$t$-normal multi-level reduced-rank FPCA model

where recall that $\boldsymbol{V}_i$ is given in (2). Using the standard result given in Anderson (1958) we can then write

$$\begin{bmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta}_i \end{bmatrix} \Big| \boldsymbol{y}_i \sim MVN\Bigg( \begin{bmatrix} \boldsymbol{D}_\alpha \boldsymbol{\Theta}_\alpha^T \boldsymbol{B}_i^T \\ \widetilde{\boldsymbol{D}_{\beta_i}} \widetilde{\boldsymbol{\Theta}_{\beta_i}^T} \widetilde{\boldsymbol{B}_i^T} \end{bmatrix} \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu),$$
$$\begin{bmatrix} \boldsymbol{D}_{\alpha_i} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{\beta_i} \end{bmatrix} - \begin{bmatrix} \boldsymbol{D}_\alpha \boldsymbol{\Theta}_\alpha^T \boldsymbol{B}_i^T \\ \widetilde{\boldsymbol{D}_{\beta_i}} \widetilde{\boldsymbol{\Theta}_{\beta_i}^T} \widetilde{\boldsymbol{B}_i^T} \end{bmatrix} \boldsymbol{V}_i^{-1} \begin{bmatrix} \boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{D}_\alpha & \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widetilde{\boldsymbol{D}_{\beta_i}} \end{bmatrix} \Bigg)$$
$$\boldsymbol{\epsilon}_i|\boldsymbol{y}_i \sim MVN\Big(\sigma^2\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu), \sigma^2\boldsymbol{I}_{N_i\times N_i} - \sigma^4\boldsymbol{V}_i^{-1}\Big)$$

Hence we immediately have

$$E[\boldsymbol{\alpha}_i|\boldsymbol{y}_i] = \widehat{\boldsymbol{\alpha}_i} = \boldsymbol{D}_\alpha\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu) \tag{10}$$
$$E[\boldsymbol{\beta}_i|\boldsymbol{y}_i] = \widehat{\boldsymbol{\beta}_i} = \widetilde{\boldsymbol{D}_{\beta_i}}\widetilde{\boldsymbol{\Theta}_{\beta_i}^T}\widetilde{\boldsymbol{B}_i^T}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{B}_i\boldsymbol{\theta}_\mu) \tag{11}$$

and the remaining results follow from the definition of covariance

$$E[\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T|\boldsymbol{y}_i] = \widehat{\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^T} = \widehat{\boldsymbol{\alpha}_i}\widehat{\boldsymbol{\alpha}_i}^T + \boldsymbol{D}_\alpha - \boldsymbol{D}_\alpha\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{D}_\alpha \tag{12}$$
$$E[\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T|\boldsymbol{y}_i] = \widehat{\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T} = \widehat{\boldsymbol{\beta}}_i\widehat{\boldsymbol{\beta}}_i^T + \widetilde{\boldsymbol{D}_{\beta_i}} - \widetilde{\boldsymbol{D}_{\beta_i}}\widetilde{\boldsymbol{\Theta}_{\beta_i}^T}\widetilde{\boldsymbol{B}_i^T}\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widetilde{\boldsymbol{D}_{\beta_i}} \tag{13}$$
$$E[\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T|\boldsymbol{y}_i] = \widehat{\boldsymbol{\alpha}_i\boldsymbol{\beta}_i^T} = \widehat{\boldsymbol{\alpha}_i}\widehat{\boldsymbol{\beta}_i}^T - \boldsymbol{D}_\alpha\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widetilde{\boldsymbol{D}_{\beta_i}} \tag{14}$$

For the case of $E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i|\boldsymbol{y}_i]$ first note that $E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i|\boldsymbol{y}_i] = tr(E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T|\boldsymbol{y}_i])$. Then, letting $\widehat{\boldsymbol{\epsilon}_i} = \boldsymbol{y}_i - \boldsymbol{B}_i\widehat{\boldsymbol{\theta}_\mu} - \boldsymbol{B}_i\widehat{\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i} - \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widehat{\boldsymbol{\beta}_i}$ and from the definition of covariance we have

$$E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T|\boldsymbol{y}_i] = \widehat{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T} = \widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^T + \sigma_i^2\boldsymbol{I}_{N_i\times N_i} - \sigma_i^4\boldsymbol{V}_i^{-1}$$

so that

$$\begin{aligned} E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i|\boldsymbol{y}_i] = \widehat{\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i} &= \text{trace}(\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^T + \sigma_i^2\boldsymbol{I}_{N_i\times N_i} - \sigma_i^4\boldsymbol{V}_i^{-1}) \\ &= \widehat{\boldsymbol{\epsilon}}_i^T\widehat{\boldsymbol{\epsilon}}_i + N_i\sigma_i^2 - \sigma_i^4 tr(\boldsymbol{V}_i^{-1}) \end{aligned} \tag{15}$$

## 2 Deriving the posterior distributions of $\gamma_{ik}$ and $\tau_{ik}$

Recall that the skew-$t$-normal density is given by

$$f(\alpha_{ik}|\xi_{\alpha_k},\sigma^2_{\alpha_k},\lambda_{\alpha_k},\nu_{\alpha_k}) = 2t_{\nu_{\alpha_k}}(\alpha_k;\xi_{\alpha_k},\sigma^2_{\alpha_k})\Phi\left(\frac{\alpha_{ik}-\xi_{\alpha_k}}{\sigma_{\alpha_k}}\lambda_{\alpha_k}\right) \tag{16}$$

The joint density $f(\alpha_{ik},\gamma_{ik},\tau_{ik})$ is given by

$$\begin{aligned} f(\alpha_{ik},\gamma_{ik},\tau_{ik}) &= f(\alpha_{ik}|\gamma_{ik},\tau_{ik})f(\gamma_{ik}|\tau_{ik})f(\tau_{ik}) \\ &= \frac{\sqrt{\tau_{ik}+\lambda^2_{\alpha_k}}}{\sqrt{2\pi}\sigma_{\alpha_k}}\exp\left\{-\frac{\tau_{ik}+\lambda^2_{\alpha_k}}{2\sigma^2_{\alpha_k}}\left(\alpha_{ik}-\xi_{\alpha_k}-\frac{\sigma_{\alpha_k}\lambda_{\alpha_k}}{\tau_{ik}+\lambda^2_{\alpha_k}}\gamma_{\alpha_k}\right)^2\right\}\times \\ &\quad I(\gamma_{ik}>0)\frac{2\sqrt{\tau_{ik}}}{\sqrt{2\pi}\sqrt{\tau_{ik}+\lambda^2_{\alpha_k}}}\exp\left\{-\frac{\tau_{ik}}{2(\tau_{ik}+\lambda^2_{\alpha_k})}\gamma^2_{ik}\right\}\times \\ &\quad \left[\frac{\nu_{\alpha_k}}{2}\right]^{\nu_{\alpha_k}/2}\tau_{ik}^{\nu_{\alpha_k}/2-1}\frac{\exp\left\{-\frac{\nu_{\alpha_k}}{2}\tau_{ik}\right\}}{\Gamma\left(\frac{\nu_{\alpha_k}}{2}\right)} \end{aligned} \tag{17}$$

We first integrate $\gamma_{ik}$ out of the joint density $f(\alpha_{ik},\tau_{ik},\gamma_{ik})$ to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} f(\alpha_{ik},\gamma_{ik},\tau_{ik})d\gamma_{ik} &= \frac{1}{\pi\sigma_{\alpha_k}}\frac{(\nu_{\alpha_k}/2)^{\nu_{\alpha_k}/2}}{\Gamma(\nu_{\alpha_k}/2)}\tau_{ik}^{(\nu_{\alpha_k}+1)/2-1}\exp\left\{-\frac{\nu_{\alpha_k}}{2}\tau_{ik}\right\}\times \\ &\quad \exp\left\{-\frac{1}{2}\frac{\tau_{ik}}{\sigma^2_{\alpha_k}}(\alpha_{ik}-\xi_{\alpha_k})^2\right\}\times \\ &\quad \int_{-\infty}^{\infty} I(\gamma_{ik}>0)\exp\left\{-\frac{1}{2}\left(\gamma_{ik}-(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)^2\right\}d\gamma_{ik} \\ &= \frac{1}{\pi\sigma_{\alpha_k}}\frac{(\nu_{\alpha_k}/2)^{\nu_{\alpha_k}/2}}{\Gamma(\nu_{\alpha_k}/2)}\tau_{ik}^{(\nu_{\alpha_k}+1)/2-1}\exp\left\{-\frac{\nu_{\alpha_k}}{2}\tau_{ik}\right\}\exp\left\{-\frac{1}{2}\frac{\tau_{ik}}{\sigma^2_{\alpha_k}}(\alpha_{ik}-\xi_{\alpha_k})^2\right\}\times \\ &\quad \int_{-\infty}^{0}\exp\left\{-\frac{1}{2}\left(\gamma_{ik}-(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)^2\right\}d\gamma_{ik} \end{aligned}$$

Making the substitution $x = \gamma_{ik}-(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}$, the integration becomes

$$\begin{aligned} &\int_{-\infty}^{0}\exp\left\{-\frac{1}{2}\left(\gamma_{ik}-(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)^2\right\}d\gamma_{ik} = \\ &\qquad \int_{-\infty}^{(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}}\exp\left\{-\frac{1}{2}x^2\right\}dx = \sqrt{2\pi}\Phi\left((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right) \end{aligned}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. Hence we have

$$\begin{aligned} f(\alpha_{ik},\tau_{ik}) &= \left(\frac{2}{\pi\sigma^2_{\alpha_k}}\right)^{1/2}\frac{(\nu_{\alpha_k}/2)^{\nu_{\alpha_k}/2}}{\Gamma(\nu_{\alpha_k}/2)}\tau_{ik}^{(\nu_{\alpha_k}+1)/2-1}\exp\left\{-\frac{\nu_{\alpha_k}}{2}\tau_{ik}\right\}\times \\ &\quad \exp\left\{-\frac{1}{2}\frac{\tau_{ik}}{\sigma^2_{\alpha_k}}(\alpha_{ik}-\xi_{\alpha_k})^2\right\}\times\Phi\left((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right) \end{aligned} \tag{18}$$

Dividing (17) by (18) gives $f(\gamma_{ik}|\alpha_{ik},\tau_{ik})$ as

$$f(\gamma_{ik}|\alpha_{ik},\tau_{ik}) = \frac{I(\gamma_{ik}>0)}{\Phi\left((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\gamma_{ik}-(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)^2\right\}$$

As $\tau_{ik}$ does not appear anywhere on the right hand side, it implies that, conditional on $\alpha_{ik}$, $\gamma_{ik}$ and $\tau_{ik}$ are independent. Hence

$$f(\gamma_{ik}|\alpha_{ik}) = \frac{I(\gamma_{ik}>0)}{\Phi\left((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\gamma_{ik}-(\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)^2\right\} \tag{19}$$

From (19) it follows that $\gamma_{ik}|\alpha_{ik} \sim TN((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}},1;(0,\infty))$ and so

$$E[\gamma_{ik}|\alpha_{ik}] = (\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}} + \frac{\phi\left((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)}{\Phi\left((\alpha_{ik}-\xi_{\alpha_k})\frac{\lambda_{\alpha_k}}{\sigma_{\alpha_k}}\right)} \tag{20}$$

For $f(\tau_{ik}|\alpha_{ik})$, we divide (18) by (16) yielding

$$f(\tau_{ik}|\alpha_{ik}) = \frac{1}{\Gamma((\nu_{\alpha_k}+1)/2)}\left(\frac{\nu_{\alpha_k}+(\alpha_{ik}-\xi_{\alpha_k})^2}{2\sigma^2_{\alpha_k}}\right)^{(\nu_{\alpha_k}+1)/2}\tau_{ik}^{(\nu_{\alpha_k}+1)/2-1}\times \exp\left\{-\frac{\tau_{ik}\left[\nu_{\alpha_k}+(\alpha_{ik}-\xi_{\alpha_k})^2\right]}{2\sigma^2_{\alpha_k}}\right\} \tag{21}$$

and so

$$\tau_{ik}|\alpha_{ik} \sim \Gamma\left(\frac{\nu_{\alpha_k}+1}{2}, \frac{\nu_{\alpha_k}+(\alpha_{ik}-\xi_{\alpha_k})^2/\sigma^2_{\alpha_k}}{2}\right)$$

# 3 Distribution of $f(\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i|\boldsymbol{y}_i,\boldsymbol{\tau}_i,\boldsymbol{\gamma}_i)$

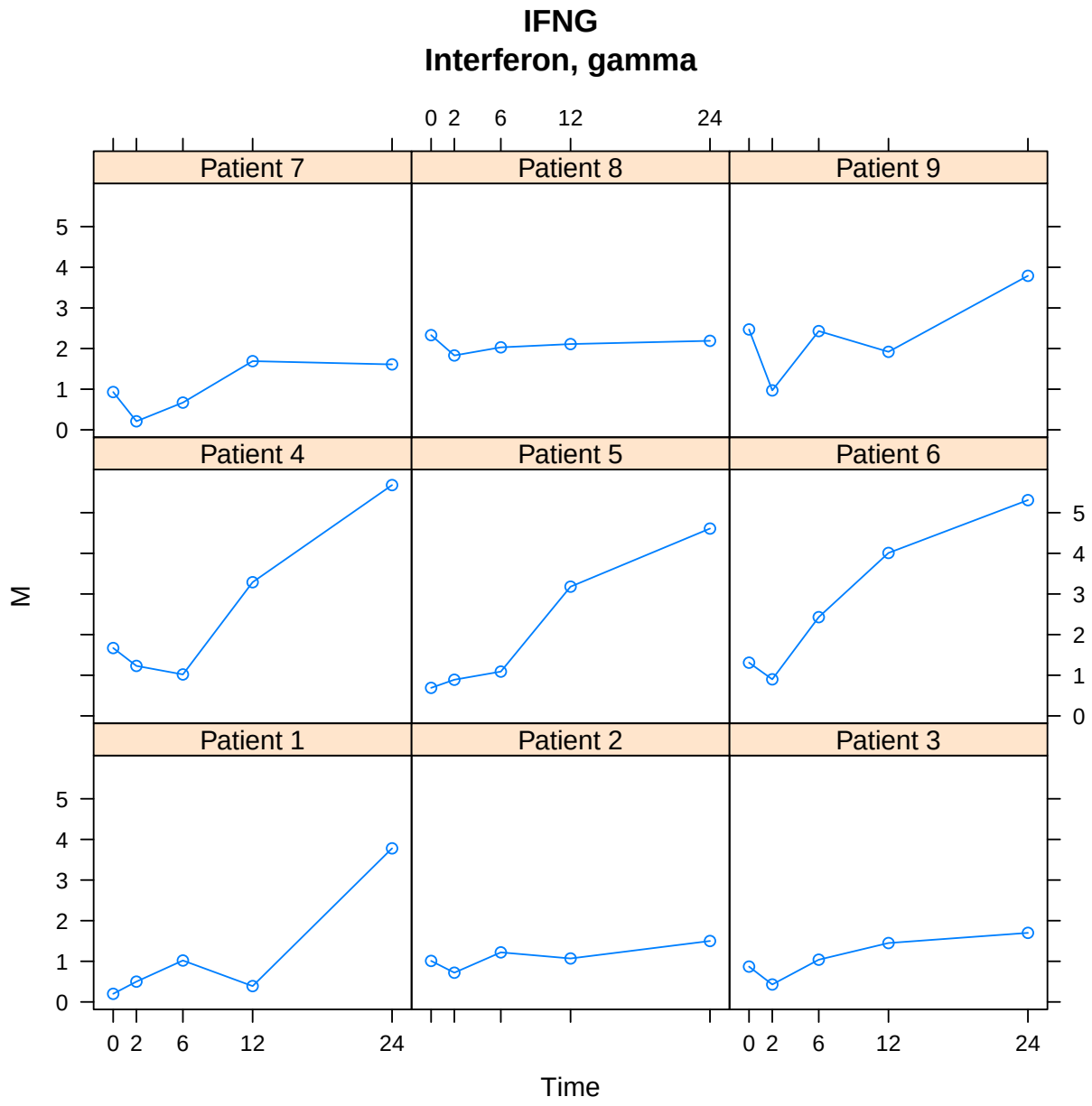
$$\begin{bmatrix}\boldsymbol{\alpha}_i\\ \boldsymbol{\beta}_i\end{bmatrix}\Big|\boldsymbol{y}_i,\boldsymbol{\tau}_i,\boldsymbol{\gamma}_i \sim MVN\Bigg(\begin{bmatrix}\boldsymbol{\mu}_\alpha\\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathrm{diag}(\boldsymbol{v}_\alpha)\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T\\ \widetilde{\boldsymbol{D}_{\beta_i}}\widetilde{\boldsymbol{\Theta}_{\beta_i}^T}\widetilde{\boldsymbol{B}_i^T}\end{bmatrix}\boldsymbol{V}^{-1}_{y_i|\tau_i,\gamma_i}(\boldsymbol{y}_i-\boldsymbol{B}_i\boldsymbol{\theta}_\mu-\boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\boldsymbol{\alpha}_i),$$
$$\begin{bmatrix}\mathrm{diag}(\boldsymbol{v}_\alpha) & \mathbf{0}\\ \mathbf{0} & \widetilde{\boldsymbol{D}_{\beta_i}}\end{bmatrix} -$$
$$\begin{bmatrix}\mathrm{diag}(\boldsymbol{v}_\alpha)\boldsymbol{\Theta}_\alpha^T\boldsymbol{B}_i^T\\ \widetilde{\boldsymbol{D}_{\beta_i}}\widetilde{\boldsymbol{\Theta}_{\beta_i}^T}\widetilde{\boldsymbol{B}_i^T}\end{bmatrix}\boldsymbol{V}^{-1}_{y_i|\tau_i,\gamma_i}\begin{bmatrix}\boldsymbol{B}_i\boldsymbol{\Theta}_\alpha\mathrm{diag}(\boldsymbol{v}_\alpha) & \widetilde{\boldsymbol{B}_i}\widetilde{\boldsymbol{\Theta}_{\beta_i}}\widetilde{\boldsymbol{D}_{\beta_i}}\end{bmatrix}\Bigg)$$

# 4 EM algorithm initialisation

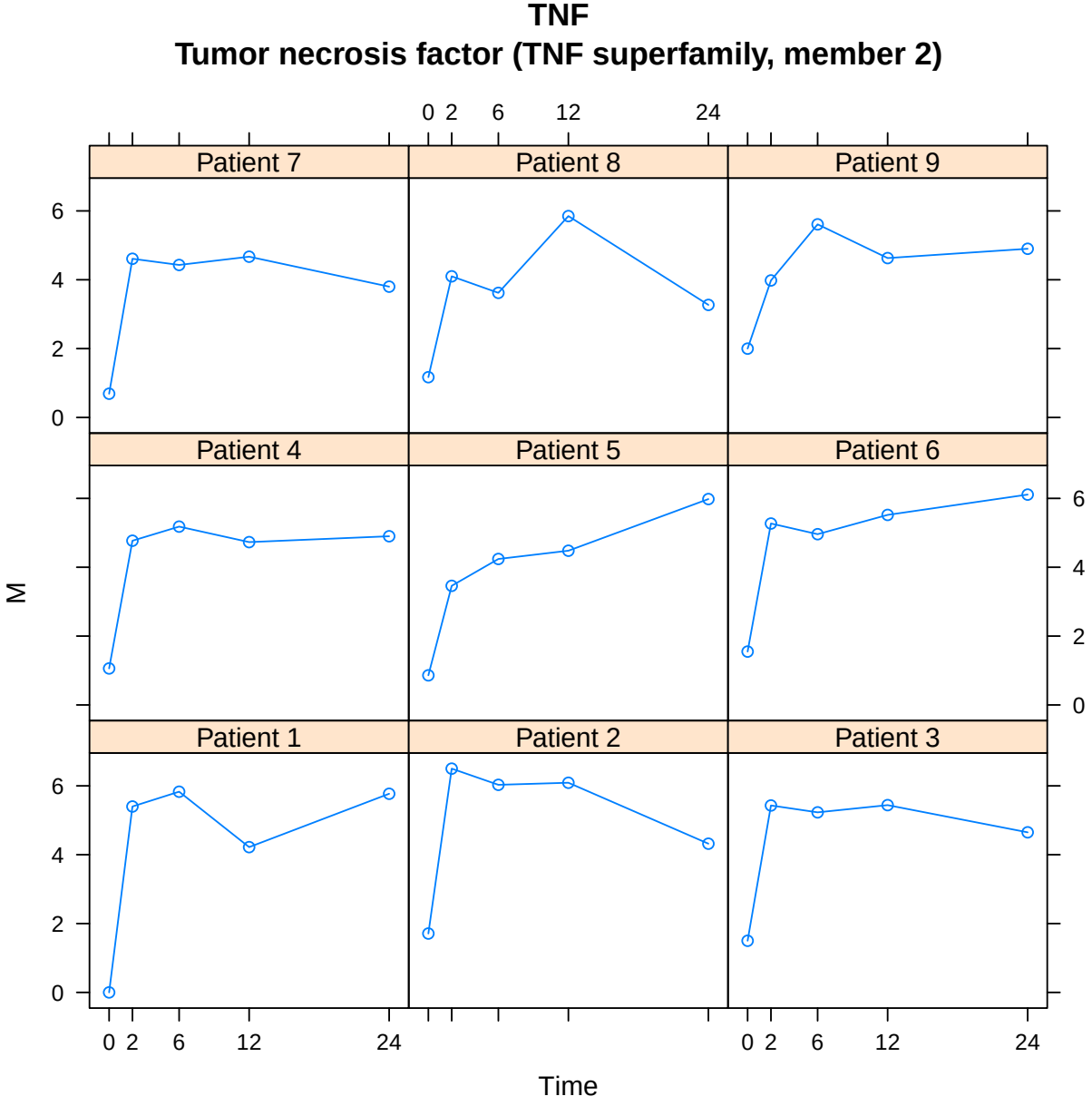
To obtain initial estimates of the parameters, the following procedure is adopted, with similar approaches used elsewhere in the literature (Peng and Paul 2009; Zhou et al. 2010).

First a spline is fit to the entire data, ignoring variable and replicate labels, using least squares in order to obtain initial values for $\widehat{\boldsymbol{\theta}_\mu}$. Next, the grand mean is subtracted from all observations and each variable

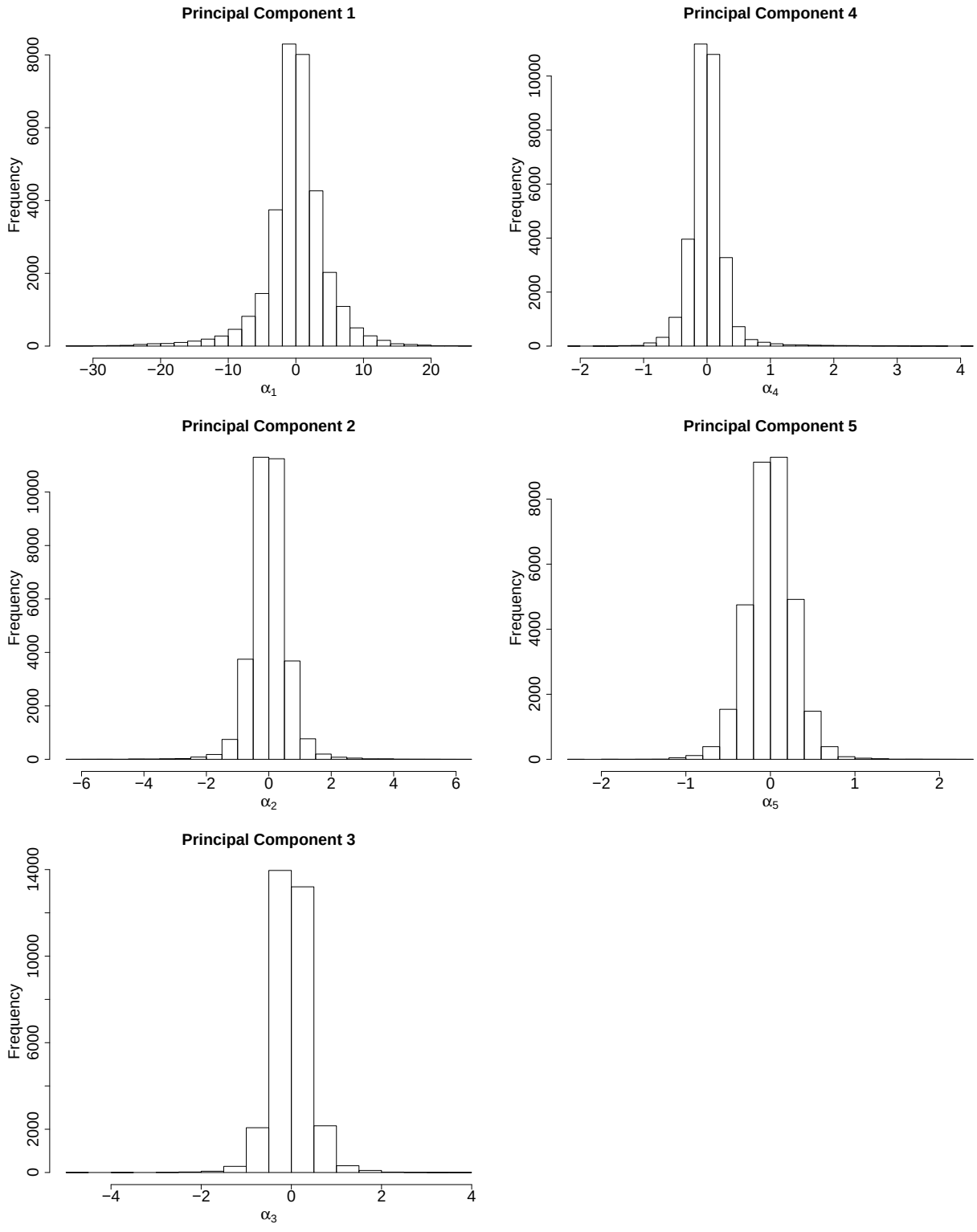
is fit independently with a spline using least squares. With $M$ variables in the data set and a $p$-dimensional spline basis, this results in an $M \times p$ matrix of spline basis coefficients. A standard PCA is performed on this matrix yielding $p$ eigenvectors each of length $p$. The first $K$ eigenvectors are taken to be the initial $\widehat{\boldsymbol{\Theta}_\alpha}$ matrix, and skew-$t$-normal distributions are fit to the loadings in order to initialise $\hat{\xi}_{\alpha_k}$, $\hat{\sigma}^2_{\alpha_k}$, $\hat{\lambda}_{\alpha_k}$ and $\hat{\nu}_{\alpha_k}$. Next the variable means are subtracted from the observations and each replicate for each variable is fit independently with a spline using least squares. Note that it will be necessary to instead perform a ridge regression in those cases where the replicate has not been observed at all time points as the matrix $\boldsymbol{B}_{ij}^T\boldsymbol{B}_{ij}$ will not be of full rank. In fact, we suggest to perform a ridge regression in all instances, even when the matrix $\boldsymbol{B}_{ij}^T\boldsymbol{B}_{ij}$ is of full rank, as this imposes some smoothness and avoids overfitting the data which may lead to unrealistically small initial values for $\widehat{\sigma_i^2}$. As a result of this procedure, an $n_i \times p$ matrix of spline basis coefficients is obtained for each variable. As before, a standard PCA is performed on this matrix and the first $L_i$ eigenvectors retained as the initial values for $\widehat{\boldsymbol{\Theta}_{\beta_i}}$. The corresponding eigenvalues are used for the initial values for the diagonal element of $\widehat{\boldsymbol{D}_{\beta_i}}$. After subtracting these replicate curves from the observations we are left with initial estimates $\widehat{\boldsymbol{\epsilon}_i}$ and the sample variance of these residuals can be used as initial values for $\widehat{\sigma_i^2}$ for each variable.



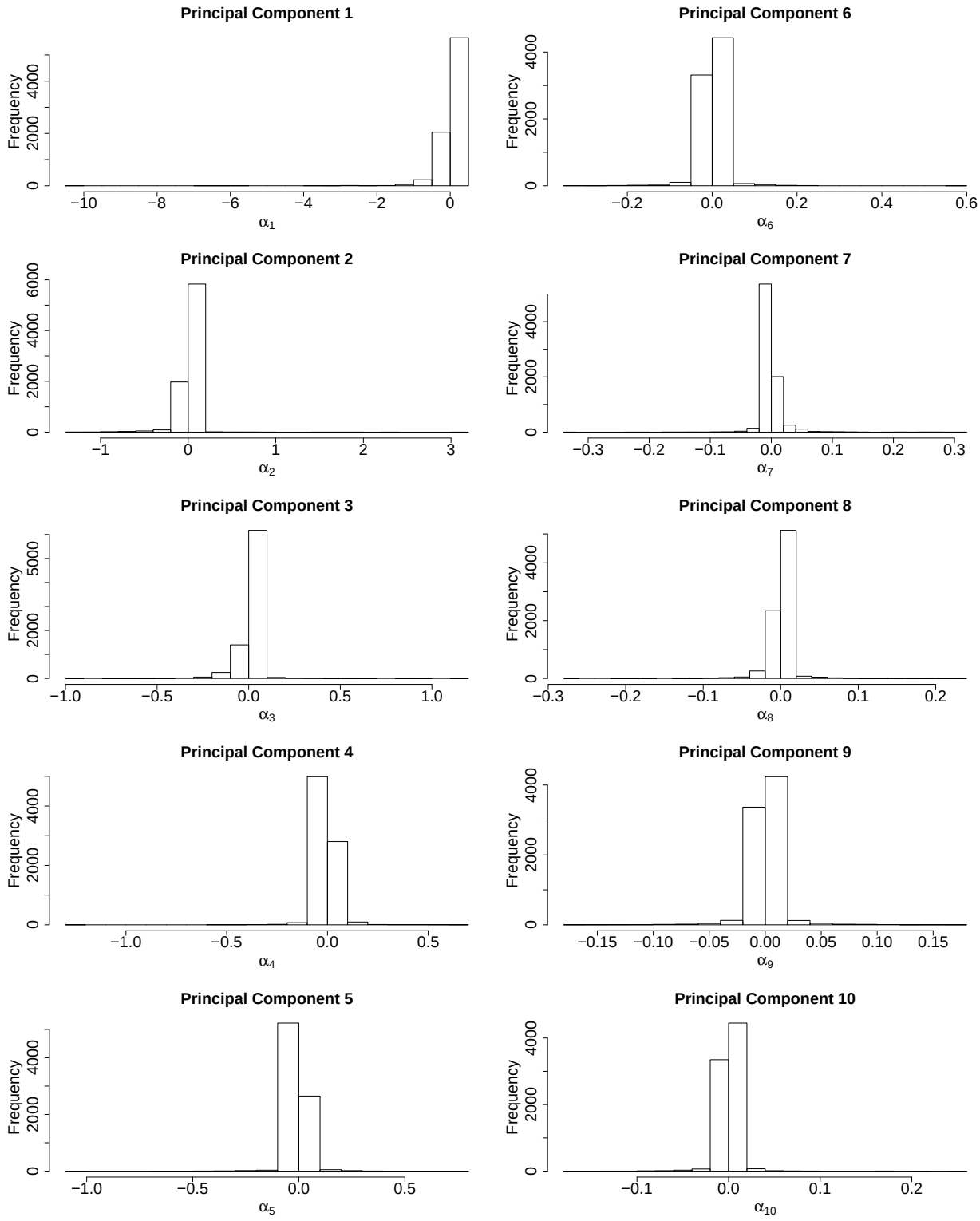

Supplementary Figure 1: Raw data for expression levels of interferon-gamma measured in blood samples from nine human patients to which the BCG vaccine for tuberculosis was added. Interferon-gamma is well-known to play an important role in the immune response to tuberculosis infection. Note the heterogeneous response with patients two, three, seven and eight exhibiting flat profiles, while the other patients end the time course with elevated levels of gene expression. Compare this to the much more homogeneous profiles for tumour necrosis factor-alpha given in Supplementary Figure 2



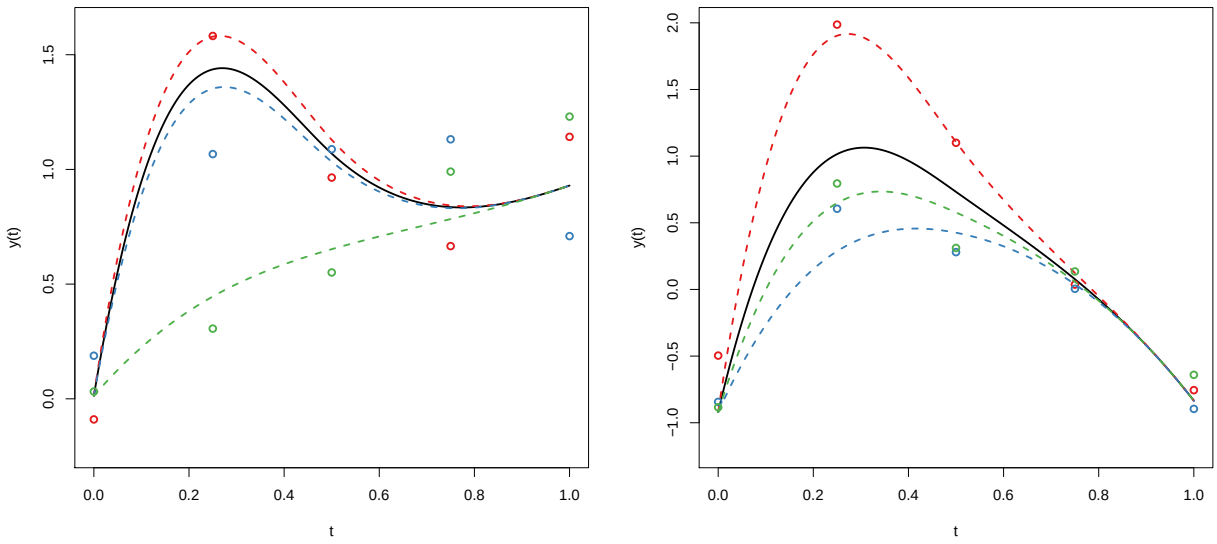

Supplementary Figure 2: Raw data for expression levels of tumour necrosis factor-alpha measured in blood samples from nine human patients to which the BCG vaccine for tuberculosis was added. Tumour necrosis factor-alpha is well-known to be associated with tuberculosis infection. Note the subtle differences between the profiles such as the expression level at which the time course begins, the level to which they peak, and whether they plateau, continue to rise or start to fall by the end of the time course. However, compared with the data given in Supplementary Figure 1 for interferon-gamma, these profiles are much more homogeneous across the different patients.



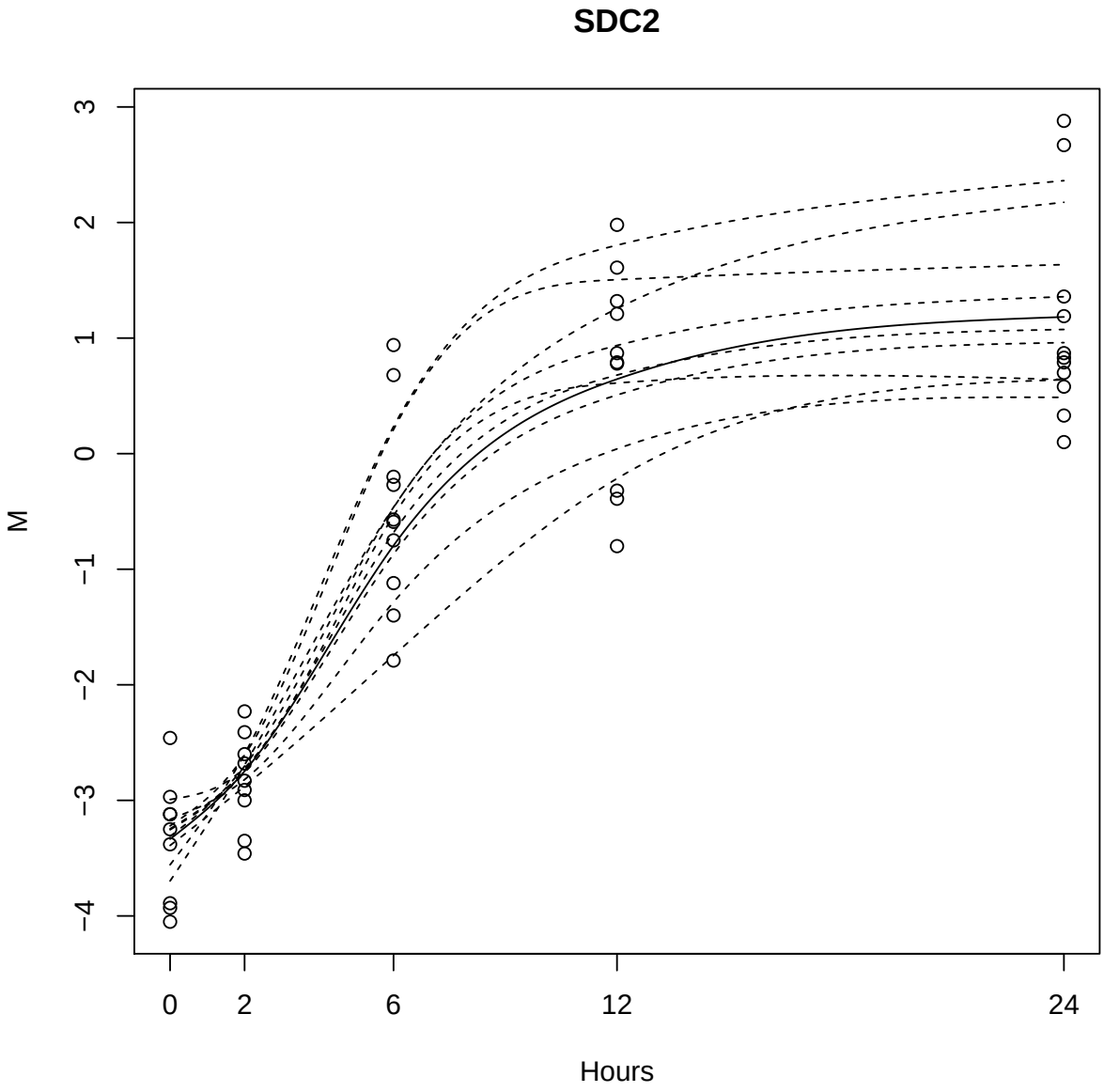

Supplementary Figure 3: Initialised variable-level loadings for a genomics data set studying the genetic response to BCG infection. Note the departure from normality in all instances with many outliers and varying levels of skewness.




Supplementary Figure 4: Initialised variable-level loadings from a metabolomics toxicology study. Note the departure from normality in all instances with many outliers and varying levels of skewness. The extreme skew for the loadings on the first principal component, such that the distribution is essentially a truncated-$t$ distribution, is a result of the fact that there are no negative observations in this data set.




Supplementary Figure 5: Example simulated data for two variables produced under the simulation setting for the Gaussian multi-level reduced-rank fPCA model. The solid black lines correspond to the variable mean curve. The coloured dashed lines are the individual replicate curves. For clarity only three replicates are shown. Final observations including simulated noise are shown as circles, and are also colour-coded for replicate. Note how the combination of the two variable-level and the single replicate-level principal component functions described above result in a wide range of replicate curves.




Supplementary Figure 6: Fit to an example transcript from the BCG genomics data set obtained under the skew-$t$-normal multi-level reduced-rank FPCA model. This transcript, corresponding to gene SDC2, was found to have the highest positive loading on the third principal component function given in Figure 5 in the main text. This principal component function account for those probes which are induced or repressed slowly until around 8 hours before levelling off. Unsurprisingly, SDC2 is a prime example of this.

| | | Number of replicates | | |
|---|---|---|---|---|
| | | 5 | 10 | 20 |
| Number of variables | 100 | 0.000406 (0.000424) | 0.000195 (0.000205) | 0.0000953 (0.0000999) |
| | 1000 | 0.000348 (0.000388) | 0.000167 (0.000186) | 0.0000813 (0.0000903) |
| | 10000 | 0.000342 (0.000387) | 0.000165 (0.000185) | 0.0000799 (0.0000897) |

Supplementary Table 3: Estimation error for variable-level curves using the Gaussian multi-level reduced-rank fPCA model. Values shown are the mean (standard deviation) MSE across all variables in 1000 simulated data sets



| | | Number of replicates | | |
|---|---|---|---|---|
| | | 5 | 10 | 20 |
| Number of variables | 100 | 0.00226 (0.00224) | 0.00113 (0.00112) | 0.000562 (0.000559) |
| | 1000 | 0.00225 (0.00223) | 0.00113 (0.00112) | 0.000564 (0.000561) |
| | 10000 | 0.00226 (0.00224) | 0.00113 (0.00112) | 0.000564 (0.000560) |

Supplementary Table 4: Estimation error for variable-level curves using the Gaussian single-level reduced-rank fPCA model. Values shown are the mean (standard deviation) MSE across all variables in 1000 simulated data sets

# References


T. W. Anderson. *Introduction to Multivariate Statistical Analysis*. Wiley, 1958.

G. James, T. Hastie, and C. Sugar. Principal component models for sparse functional data. *Biometrika*, 87 (3):587–602, 2000.

J. Peng and D. Paul. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18(4): 995–1015, 2009.

L. Zhou, J. Z. Huang, J. G. Martinez, A. Maity, V. Baladandayuthapani, and R. J. Carroll. Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, 105(489):390–400, 2010.